TEC-0057

Representation, Modeling and Recognition of Outdoor Scenes Second Annual Report

Martin A. Fischler
Robert C. Bolles

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493

April 1994

Prepared for:
Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203-1714

Monitored by:
U.S. Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Road
Alexandria, Virginia 22315-3864

US Army Corps of Engineers
Topographic Engineering Center

T

E

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

The citation in this report of trade names of commercially available products does not constitute official endorsement or approval of the use of such products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1994	Second Annual Report Apr. 1993 - Apr. 1994

4. TITLE AND SUBTITLE

Representation, Modeling and Recognition of Outdoor Scenes
Second Annual Report

5. FUNDING NUMBERS

DACA76-92-C-0008

6. AUTHOR(S)

Martin A. Fischler Robert C. Bolles

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Advanced Research Projects Agency
3701 North Fairfax Drive, Arlington, VA 22203-1714

U.S. Army Topographic Engineering Center
7701 Telegraph Road., Alexandria, VA 22315-3864

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

TEC-0057

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The primary goal of this project was to advance the state-of-the-art in scene interpretation for autonomous systems that operate in natural terrain. In particular, techniques are being developed for representing knowledge about complex cultural and natural environments so that a computer vision system can successfully plan, navigate, recognize and manipulate objects and answer questions or make decisions relevant to this knowledge. The results include the development of representations and associated methods for rapidly modeling natural terrain, from multiple images, at a level or organization higher than that of the conventional dense array of depths. In particular, we have developed an approach for integration of information acquired from multiple views of a scene that uses a new class of geometric primitives; it allows easy expression of known constraints and observed data, and also allows the use of practical optimization based solution techniques. This work will provide an effective way of allowing a robotic system to incrementally build a progressively more accurate and complete model of the environment in which it is operating. Work is also progressing on the problem of recognizing important classes of natural and man-made objects -- especially roads, trees, rocks, and terrain features.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Machine Vision, Automated Scene Analysis, Object Recognition, Terrain Modeling	57
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

TABLE OF CONTENTS

PREFACE

This research is sponsored by the Advanced Research Projects Agency (ARPA) and monitored by the U.S. Army Topographic Engineering Center (TEC) under Contract DACA76-92-C-0008, titled "Representation, Modeling and Recognition of Outdoor Scenes, Second Annual Report". The ARPA Program Manager is Mr. Charles Shoemaker, and the TEC Contracting Officer's Representative is Ms. Lauretta Williams.

1. OBJECTIVE

Our primary goal in this project is to advance the state of the art in scene interpretation for autonomous systems that operate in natural terrain. In particular, techniques are being developed for representing knowledge about complex cultural and natural environments so that a computer vision system can successfully plan, navigate, recognize, and manipulate objects, and answer questions or make decisions relevant to this knowledge.

2. APPROACH

This work integrates advances in four separate technologies to achieve the goal of providing a foundation for the design of highly competent machine vision systems capable of autonomous operation in the outdoor world.

First, stored knowledge (such as map data and object models) provides the basis for invoking context, function, and purpose, in addition to the use of visually observed geometric shape, to recognize scene objects.

Second, we are developing compact and expressive representations for modeling, and ultimately recognizing, objects encountered in the natural world. Computational efficiency, and thus real time performance, is critically dependent on using effective representations for both models and sensed data.

Third, global optimization techniques are being developed that require reasonable amounts of computation, but which are expected to produce results beyond those obtainable by local analysis methods.

Fourth, techniques are being developed that are able to simultaneously, or incrementally, exploit multiple views of a scene in compiling a complete scene model. For example, in our previous work we have been able to demonstrate that the integrated analysis of a motion sequence can be used to construct a geometric scene model that is superior to a sequence of independent stereo reconstructions.

3. PROGRESS

This program is a continuous long-term effort addressing important but difficult problems; it builds on our previous Advanced Research Projects Agency (ARPA) research. Our results to-date are centered on the development of representations and associated methods for rapidly modeling natural terrain (from multiple images) at a level of organization higher than that of the conventional dense array of depths. This work will provide the essential advance needed to turn raw geometric measurements into timely information usable by robotic navigation and planning systems. Work is also progressing on two additional problems: modeling compact 3-D objects from their projected 2-D contours, and the problem of recognizing important classes of natural and man-made objects -- especially roads, trees, rocks, and terrain features.

4. SUMMARY OF RECENT ACCOMPLISHMENTS

We continue the development of an approach for integration of information acquired from multiple views of a scene into a description of scene geometry. The approach uses a new class of geometric primitives which allows easy expression of known constraints and observed data, and also allows the use of practical optimization based solution techniques. This work will provide an effective way of allowing a robotic system to incrementally build a progressively more accurate and complete model of the

environment in which it is operating. Papers describing this work have been accepted for journal publication (IJCV) and conference presentation (SPIE, CVPR, and ECCV). Two new papers (included as appendices in this report), just completed, describe extensions of our work in this area to more general imaging situations and can handle additional constraints. A more detailed discussion of this topic area is provided in a following section.

The problem of automatically recognizing objects appearing in images of the outdoor world has proven to be extremely difficult, in part, because in addition to all the other difficulties of object recognition, we must now also contend with the lack of explicit shape models. While most of the current (successful) computer-based recognition approaches rely on explicit knowledge of shape, rocks, trees, and other natural objects cannot be successfully described in this way; even such generic man-made objects as roads, bridges, and buildings are more likely to satisfy functional constraints rather than being exemplars of some geometric blueprint. In order to replace explicit shape with a more general way of describing natural objects (and complex man-made structures), a large number of geometric primitives have been proposed that are also suitable for detection by automatic image analysis algorithms (e.g., edges, textures, fractals). The result of much of this past work is that, while often promising, the techniques are not sufficiently reliable to provide a basis for the knowledge-based analysis needed to complete the recognition task. What is required are a few techniques than can very reliably organize the pixel-level image data as a basis for higher level analysis. Finding the appropriate combination of low-level data-description, and associated extraction techniques, is thus a key problem in machine vision and one of our primary concerns in this project. In addition to our work relevant to this topic discussed above, we have focused on extracting coherent line (as distinct from edge) features in single gray-level images. We note that a line sketch of some object or scene is often sufficient to depict the imaged information in a very compact way.

Two techniques have emerged from this work that appear to meet the criterion of generality and robustness. The first is a generic way to find candidate line structure in an image; some of this work is described below. The second is a way to organize such data into perceptually coherent and semantically meaningful units. In a recently completed paper (IEEE-PAMI 1994) we describe our progress in the design of a curve partitioning technique that is extremely robust in achieving the perceptual organization task; we also describe how this technique can be applied to the problem of road delineation in aerial images.

In our most recent work in this general subject area, we are focusing on the development of algorithms for the recognition of a number of natural objects of importance to robotic navigation and outdoor scene modeling -- and in developing metrics for objective evaluation of scientific progress (we have taken a lead position in ARPA's benchmarking efforts in natural object recognition). Some of this work is described below and a paper is in preparation.

Earlier in this project, we made a significant new advance in the long-standing problem of duplicating human performance in recovering 3-D models of terrain and man-made objects from qualitative and imprecise line drawings (e.g., of terrain elevations as in an approximate and uncalibrated contour map, or building edges as in a single approximate projection of the corresponding wire-frame). This work can greatly simplify communication problems between man and machine in such applications as robotic

mission planning and in construction of databases for use in robotic navigation. A paper describing this work has been published ("An optimization based approach to the interpretation of single line drawings as 3-D wire frames," IJCV 9(2):113-136, November 1992). On-going work has led to (new) additional results of both theoretical and practical importance; these new results, still being further developed and evaluated, will be described in a later report.

5. DETAILED DISCUSSION OF WORK ON GEOMETRIC RECONSTRUCTION FROM MULTIPLE VIEWS

To reconstruct object surfaces, one can start with a number of measuring techniques, for example, laser rangefinding, stereo or 3D scanners, all of which provide raw information about the location of points in space. These points, however, are often from noisy "clouds" of data instead of the surfaces one expects.

Deriving the surfaces from such data is a difficult task because:

- The 3D points may form a very irregular sampling of the space,
- They may have been produced by several sensors or derived from several viewpoints so that it becomes impossible to work only in the imaging plane of any one sensor,
- Several surfaces can overlap; simple interpolation will not work,
- The sensors and algorithms make mistakes that must be properly dealt with.

In this research effort, we address the problem of determining the 3-D shape and material properties of surfaces by combining the information provided by active or passive ranging techniques with that present in multiple 2-D intensity images. We have developed an optimization-based surface reconstruction method that relies on an object-centered representation to recover 3-D surfaces. Our method uses both monocular shading cues and stereoscopic cues from any number of images, while correctly handling self-occlusions. It can also take advantage of the geometric constraints derived from measured 3--D points and 2--D silhouettes. These complementary sources of information are combined in a unified manner so that new ones can be added easily as they become available. Using a variety of real imagery, we have demonstrated that the resulting method is quite powerful and flexible, allowing for both completely automatic reconstruction in straightforward circumstances, and for user-assisted reconstruction in more complex ones. User assistance is so far provided primarily through the introduction of a small number of hand-entered linear and point features using semiautomated "snake" technology. The method is controlled by a small number of image-independent parameters that specify the relative importance of the various information sources.

(As discussed in previous reports, we are also investigating a "local" approach to surface recovery; this older approach has some important aspects that are not currently available in the global recovery technique. We intend to merge the two methods and will present the details of this synthesis if it proves successful.)

5.1 Surface Geometry

Given camera models for the images being analyzed, the corresponding projections of the 3-D surface points appearing in the images can be computed, and making the usual stereo assumption, must have comparable grey levels. Our algorithm optimizes the placement of surface vertices to minimize the overall difference in grey levels while preserving surface smoothness. The actual criterion we use is a linear combination of the sums of the variance of grey levels across images and of the sums of the surface curvatures at the

vertices. We use a conjugate-gradient, and (recently) other descent algorithms, embedded in a continuation method to perform the optimization: we first optimize with a strong smoothness constraint; we then reduce the constraint progressively.
Because our surfaces are 3-D objects, we can directly determine the presence of hidden surfaces and deal effectively with occlusions. In order to detect those hidden surfaces in an effective manner, we have implemented the algorithm to run on an SGI machine and exploit the machines z-buffering capabilities.

In most of our past experiments, we have used regular grids and uniform smoothness constraints. While this is appropriate for surfaces whose properties remain relatively constant, this is suboptimal for more complex surfaces that can be more effectively handled using triangulated irregular networks. The relatively smooth parts of such surfaces should be represented by large patches while the rougher parts are better described by finer and less constrained triangulations. We have made progress in implementing such irregular networks by allowing some of the regular facets to be subdivided as required by the surface geometry.

In particular, in our integration of raw depth data to produce higher level terrain descriptions (for tasks such as obstacle detection), we are faced with the problem of how to retain geometric accuracy once the point arrays of scene depths are no longer available. We have developed an algorithm for adaptively choosing minimal triangulations with fixed error bounds. Beginning with a high resolution regular triangulation, we simultaneously minimize a robust measure of error-of-fit and reduce the number of triangles in a local region whenever such a reduction does not violate the accuracy constraint. This constraint is instantiated as a fixed envelope surrounding the data, within which the triangulation surface must always fall. Since the raw depth data is accurate to sub-pixel disparities and the triangulation error is explicitly bounded everywhere, the resulting surface is no less accurate, yet it is a minimal description of the disparity surface.

We have also begun developing an algorithm that segments potential obstacles from the ground-level surface and models the latter. The algorithm starts with a disparity map and the cloud of 3--D points it implies. These points are treated as attractors that generate a potential field in which we embed an initially planar model of the ground-level surface. We then deform this surface model to minimize its potential energy. Next we compute the distance of the attractors to the surface, discount those that are farthest from it, reoptimize, and then repeat the whole process. After a few iterations, the ground surface model stabilizes and yields a description of both the ground geometry and the potential obstacles, i.e. the points that appear to be off the ground.

5.2 Physical Properties

Many natural surfaces can be modeled by a Lambertian reflectance model whose albedo depends on the corresponding physical surface properties. Recovering this albedo is therefore an important first step towards the goal of analyzing those physical properties and potentially segmenting regions of interest. Unlike traditional "shape from shading" approaches that work in image space and assume constant albedo, our technique allows us to assign different albedoes to the facets of the derived triangulation. We can then optimize the values assigned to these albedoes and also find (or use the known) location of the light source to maximize the similarity between the shaded image derived from our models and the real images.

We are performing experiments with the above method for computing albedo given surfaces originally derived using stereo. The objective function we optimize enforces

albedo smoothness while minimizing intensity difference between the shaded images and the real ones. To make this approach fully general, we will introduce albedo discontinuities to account for abrupt changes in surface material type. We will also attempt to determine those classes of natural objects and terrain types for which the Lambertian model is appropriate by examining the variance in intensity across images of the same scene acquired from different viewpoints.

Our ultimate goal in the above two tasks is to be able to optimize simultaneously the vertex positions and the surface albedoes in order to compute surface geometry and photometry. Our current focus in this task is to combine the stereo objective function with the photometric one in order to achieve a more complete description of the scene.

5.3 Use of Additional Contextual, Photometric, and Geometric Constraints

We have added to our system the capability of using the constraints generated by 3--D points, 3--D linear features, and 2--D silhouettes. This capability is important for applications such as high-resolution cartography. One must ensure that the terrain model conforms to the feature data and does not violate any physical constraints: roads should be on the ground and not overly tilted, streams should stay within stream beds, buildings should not be floating in space, and so on. Our method allows one to both satisfy these constraints and account, as well as possible, for the observed image data.

We deal with the various sources of 3--D information, whether dense, such as range maps or correlation-based stereo disparity maps, or linear, such as hand-entered features or edge-based stereo disparity maps, in the same fashion. Both are sampled at regular intervals to generate collections of 3--D attractors that are used to define energy terms that are added to our overall objective function. (i.e., in the case of features whose 3-D coordinates can be computed, we modify the objective function that governs the behavior of the deformable surface by adding a term that attracts some of its vertices towards the features and at the same time turns off the smoothing for those vertices.)

Silhouettes, or occlusion contours, are 2--D features since they depend on the viewpoint and cannot be matched across images. However they constrain the surface tangent. Each point of the silhouette edge defines a line that goes through the optical center of the camera and is tangent to the surface at its point of contact with the surface. The points of a silhouette edge, therefore, define a ruled surface that is tangent to the surface. As previously, we take advantage of these constraints by sampling the silhouette edges and adding new terms to our objective function.

To take full advantage of these constraints, we have augmented our set of available optimization techniques by combining a "snake-like" approach and a coarse-to-fine one. In the "snake-like" approach, we embed the surface into a viscous medium and solve the dynamics equation by solving, at every iteration, a set of linear equations. This has proved more effective than the conjugate-gradient method we were using before. In the coarse-to-fine approach, we start with a large-grained mesh that we progressively refine by subdividing the facets.

Finally we can use our reconstructed surfaces to predict shadow locations and compare them to the actual location for verification purposes.

5.4 Implementation and Testing

We continue to refine and test our method for reconstructing both the shape and reflectance properties of physical surfaces from the information present in multiple images. We have, so far, considered two classes of information. The first class contains the information that can be extracted from a single image, such as texture gradients, shading, and occlusion edges. We take advantage of the fact that multiple images enhance the utility of this type of information by allowing for consistency checks across the images as well as the use of averaging to improve precision. The second class contains information that require at least two images for its extraction, such as the depth of corresponding points found in two input images through the use of stereo triangulation.

As previously discussed, our surface reconstruction method uses an object-centered representation, specifically, a hexagonally-connected 3-D mesh of vertices with triangular facets. Such a representation accommodates the two classes of information mentioned above, as well as multiple images (including motion sequences of a rigid object) and self-occlusions. We have chosen to model the surface material using the Lambertian reflectance model with variable albedo, though generalizations to specular surfaces are currently being evaluated. Consequently, the natural choice for the monocular information source is shading, while intensity is the natural choice for the image feature used in multi-image correspondence. Not only are these the natural choices when we are able to assume a Lambertian reflectance model, they are complementary: intensity correlation is most accurate wherever the input images are highly textured, and shading is most accurate when the input images have smooth intensity variation. Since we wish to deal with surfaces with non-uniform albedo, we have developed a new approach to incorporating shading information that uses the variation in computed albedo from facet to facet as the indicator of a correct surface reconstruction.

We use an optimization approach to reconstruct the surface shape and its material properties from the input images. That is, we alter the shape and reflectance properties of the surface mesh so as to minimize an objective function, given an initial surface estimate provided by other means, such as a standard stereo algorithm. The objective function is a linear combination of an intensity correlation component, an albedo variation component, and a surface smoothness component. The first two components are a function of the intensities projected onto the triangular facets from the input images (taking occlusions into account), and are weighted according to the amount of texture in the intensities, for the reasons mentioned in the previous paragraph. The geometric smoothness component is slowly decreased during the optimization process to allow for an accurate estimate of the surface shape and reflectance.

We have implemented an algorithm employing these three terms and have performed extensive experiments using synthetic images as well as real aerial and face images. The strengths of the approach include:

- The use of the 3-D surface mesh allows us to deal with self-occlusions, and thus, effectively merge information from several potentially very different viewpoints to eliminate "blind-spots."

- By combining stereo and shape from shading, and weighing appropriately the reliability of their respective contributions, we can obtain results that are better than those produced by either technique alone.

- Using the facets to perform the stereo computation frees us from the constant-depth assumption that standard correlation-based stereo techniques make. It becomes possible to recover accurately the depth of sharply sloping surfaces (such as that of a sharp ridge).

- The shape from shading component does not make a constant-albedo assumption unlike most shading algorithms. Instead, we only make the weaker and much more general assumption that albedoes vary slowly across textureless areas.

- As in some of our most recent work, we can continue to introduce additional contextual, photometric, and geometric constraints into our implemented system by simply altering the objective function and occasionally making very minor changes in the code. These additional constraints can greatly enhance the competence of the system at a small cost in computation time.

6. DETAILED DISCUSSION OF WORK ON NATURAL OBJECT RECOGNITION

In this effort we are developing and evaluating two distinct approaches to natural object recognition. The first is based primarily on geometric information about the scene as obtained from stereo or some other form of range sensing. The second approach is based on using information about illumination, shadows, shape, color, and texture as extracted from a single color image of the scene (or possibly from multiple images -- but not requiring the explicit availability of range data). If both approaches prove successful, we will also consider merging them into a combined technique.

6.1 Rock and Small Object/Obstacle Recognition (to Permit Robotic Navigation)

Because of their importance to Unmanned Ground Vehicle (UGV) navigation, techniques for the recognition and delineation of rocks and other relatively small objects and terrain features are being constructed using a number of distinct approaches. (It is intended to make these techniques sufficiently general so as to allow their use in finding a wide range of compact objects protruding above the ground surface).

In one investigation, we are concerned with locating rock-like objects and obstacles within the operating environment of a forward looking pair of stereo sensors. Here we adopt a very simple geometric model of these obstacles in order to satisfy the requirements for high speed detection and limited computing power needed for practical UGV applications.

We begin by assuming that these obstacles are distributed sparsely enough so that we may examine each in relative isolation and locate them using only simple local patterns in the disparity image(s). When the obstacles are spatially separated, the detection patterns depend primarily on the conjunction of disparity discontinuities (e.g., occlusion edges) and stereo shadows (regions only visible from one camera). We are currently conducting experiments to determine whether or not this approach is robust in the presence of multiple objects overlapping along a line-of-sight.

A key feature of the local pattern used to locate these obstacles is the "top" of the object, a near-horizontal spatial discontinuity. This feature is flanked by either a stereo shadow region or a high disparity gradient. If both L-R and R-L disparity images are made available, then mirror images of this pattern may be sought in the two disparity images.

The coordinated detection of these features locates the presence and extent of such obstacles.

In the simplest application (to obstacle detection) of the above approach, varying the thresholds throughout the image, dependent on the stereo geometry, allows the detection of obstacles based solely on size. This type of analysis depends critically on the parameters of the low-level stereo fusion used. The resolution determines the lower limit on the size of obstacles detectable. The amount of smoothing used to condition the results also affects detection, especially if the method used obscures discontinuities. In general, however, without additional texture or chromatic input, it is difficult to distinguish between obstacles (e.g. rocks) and non-obstacles (e.g., clumps of grass). In the work discussed below, we respond to the need to distinguish actual hazards, such as rocks, from benign, but similarly shaped objects, such as bushes.

6.2 Semantic Scene Description and Modeling of Extended Terrain Features

In addition to supporting local robotic navigation, we are also concerned with the more general problems of creating a semantic scene description for use by the higher level robotic planning and control systems, and for use in other applications (e.g., mapping, simulation, and automatic target recognition).

Our goal in this task is to be able to recognize the natural objects and terrain features in the context of creating an overall scene sketch. That is, we are not only interested in recognizing (and delineating) isolated objects, but wish to describe and exploit their interrelationships. Objects of interest include: rocks, trees, brush, grass, water, snow, dirt, sky, ridgelines, holes/ditches, roads, paths, fences, poles, cliffs, ground-plane, and shadows. Our approach here is to first select (or define) a smaller set of primitive (but pervasive) features that can reliably be extracted from most images of natural scenes. This set of primitives (currently consisting of: color, texture, shadows, depth, surface orientation, and linear structures) are combined to identify clear instances of the natural objects of interest using a "production rule" type paradigm (ref: Strat and Fischler, "Context-based vision," IEEE PAMI, October 1991), and then, using these recognized objects as exemplars, we invoke a nearest-neighbor statistical classifier to label other, possibly less obvious, instances of the objects we are looking for.

Thus, for example, we have implemented a color-based classification algorithm using classical feature-space partitioning techniques and run a significant number of experiments on outdoor images from a variety of sources. We have been successful in distinguishing certain categories of objects, such as between vegetation and rocks (or other non-living terrain features). While not sufficient to produce a complete semantic labeling of a scene at a detailed level of recognition, this approach appears able to complement our geometric recognition techniques and allow us to successfully deal with some of the critical open problems -- such as distinguishing between real obstacles (e.g., rocks) and navigable areas (e.g., grass or brush covered flat terrain).

We are also developing techniques to recognize the more important and prominent extended linear terrain features and navigation obstacles; these include the skyline, ridgelines, and the leading edge of drop-offs and ditches. These features appear as significant linear discontinuities in the disparity image, and as intensity discontinuities in a color or grayscale image. By first locating depth/disparity and intensity or color discontinuities, and then linking these local features, we can find and label both small compact objects and the major linear scene features. If we are successful in delineating the skyline in a color image, then we choose a region above the skyline as an exemplar of "sky" for use by our color classifier. Texture, shadows, and shape will allow us to find a

few obvious instances of vegetation and rocks as additional exemplars for the color classifier. We can now label most of the scene using the color classifier, and then, (at least partially) check the result for semantic consistency: The pixels labeled sky by the color classifier should all lie above the skyline found by the linear delineation process; if the skyline is interrupted by a nearby thin raised object, the object should be labeled as a tree or a pole; a relatively horizontal/flat region, depending on its color, should be either grass, dirt, water, or rock; etc.

As an additional part of our effort to construct an overall scene sketch, we are also developing techniques to build general geometric descriptions of the terrain, using triangulated meshes. A mesh is first generated using information provided by the disparity image, then it is reduced in complexity so as to minimize the number of triangles in the mesh. The technique developed for this is similar to "mesh decimation" used in graphics, but constant reference is maintained to the initial disparity image, thus ensuring that the final mesh is still an accurate representation of the disparity image. We have been able to achieve reductions of over 98% of the data in the original mesh. When this mesh is transformed into real-world coordinates we have an accurate, triangulated description of the visible surfaces.

Appendix A

"Using 3-Dimensional Meshes to Combine Image-Based and Geometry-Based Constraints"

ECCV; Stockholm, Sweden, May, 1994
P.V. Fua and Y.G. Leclerc

Using 3–Dimensional Meshes To Combine Image-Based and Geometry-Based Constraints

P. Fua and Y.G. Leclerc
SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025, USA
(fua@ai.sri.com leclerc@ai.sri.com)

Abstract

In this paper, we present a unified framework for 3–D shape reconstruction that allows us to combine image-based and geometry-based information sources. The image information is akin to stereo and shape-from-shading while the geometric information may be provided in the form of 3–D points, 3–D features or 2–D silhouettes. A formal integration framework is critical because, in order to recover complicated surfaces, the information from a single source is often insufficient to provide a unique answer.

Our approach to shape recovery is to deform a generic object-centered 3–D representation of the surface so as to minimize an objective function. This objective function is a weighted sum of the contributions of the various information sources. We describe these various terms individually, our weighting scheme and our optimization method. Finally, we present results on a number of difficult images of real scenes for which a single source of information would have proved insufficient.

Keywords : Surface reconstruction, Stereo, Shape-from-Shading, Silhouettes, Geometric Constraints.

1 Introduction

The problem of recovering surface shape from image cues, the so-called "shape from X" problem, has received tremendous attention in the computer vision community. But no single source of information "X," be it stereo, shading, texture, geometric constraints or any other, has proved to be sufficient across a reasonable sampling of images. To get good reconstructions of a surface, it is necessary to use as many different kinds of cues with as many views of the surface as possible. In this paper, we present and demonstrate a working framework for surface reconstruction that combines image cues, such as stereo and shape-from-shading, with geometric constraints, such as those provided by laser range finders, area- and edge-based stereo algorithms, linear features and silhouettes.

Our framework can incorporate cues from many images of a surface, even when the images are taken from widely differing viewpoints, accomodating such viewpoint-dependent effects as self-occlusion and self-shadowing. It accomplishes this by using a full 3–D object-centered representation of the estimated surface. This representation is then used to generate synthetic views of the estimated surface from the view of each input image. By using standard computer graphics algorithms, those parts of the surface that are hidden from a given viewpoint can be identified and consequently eliminated from the reconstruction process. The remaining parts are then in correspondence with the input images, and the images and corresponding cues are applied to the reconstruction of the surface in an iterative manner using an optimization algorithm.

Recent publications describe the reconstruction of a surface using 3–D object-centered representations, such as 3–D surface meshes [Cohen *et al.*, 1991, Delingette *et al.*, 1991, Terzopoulos and Vasilescu, 1991, Vemuri and Malladi, 1991], parameterized surfaces [Stokely and Wu, 1992, Lowe, 1991], local surfaces [Ferrie *et al.*, 1992, Fua and Sander, 1992], particle systems [Szeliski and Tonnesen, 1992], and volumetric models [Pentland, 1990, Terzopoulos and Metaxas, 1991, Pentland and Sclaroff, 1991]. Most of these rely on previously computed 3–D data, such as the coordinates of points derived from laser range finders or correlation-based stereo algorithms, and reconstruct the surface by fitting it to these data in a least-squares sense. In other words, the derivation of the 3–D data from the images is completely divorced from the reconstruction of the surface.

In contrast, our framework allows us to directly use such image cues as stereo, shading, and silhouette edges in the reconstruction process while simultaneously incorporating previously computed 3–D data such as those mentioned above. In a previous publication [Fua and Leclerc, 1993] we describe how stereo and shading are used within the framework described below, and the relationship of this approach to previous work. Here, we focus on how an additional image cue (silhouette edges) and previously computed 3–D data are incorporated into our reconstruction process.

Combining these different sources of information is not a new idea in itself. For example, Blake *et al.* [1985] is the earliest reference we are aware of that discusses the complementary nature of stereo and shape from shading. Both Cryer *et al.* [1992] and Heipke *et al.* [1992] have proposed algorithms to combine shape-from-shading and stereo while Liedtke *et al.* [1991] first uses silhouettes to derive an initial estimate of the surface, and then applies a multi-image stereo algorithm to improve the result. However, none of the algorithms we know of uses an object-centered representation and an optimization procedure that are general enough to incorporate all of the cues that we present here. This generality should also make possible the use of a very wide range of other sources of information, such as shadows, in addition to those actually discussed here.

We view the contribution of this paper as providing both the framework that allows us to combine diverse sources of information in a unified and computationally effective manner, and the specific details of how these diverse sources of information are derived from the images.

In the next section, we describe our framework and the new information sources introduced here. Following this, we demonstrate that the framework successfully performs its function on real images and allows us to achieve results that are better than those we could derive from any single source of information (and

often better than any two sources of information).

2 Framework

Our approach to recovering surface shape and reflectance properties from multiple images is to deform a 3-D representation of the surface so as to minimize an objective function. The free variables of this objective function are the coordinates of the vertices of the mesh representing the surface, and the process is started with an initial estimate of the surface. Here we assume that images are monochrome, and that their camera models are known *a priori*.

We represent a surface S by a hexagonally connected set of vertices $\mathbf{V} = (v_1, v_2, \ldots, v_{n_v})$ called a *mesh*. The position of vertex v_j is specified by its Cartesian coordinates (x_j, y_j, z_j). Each vertex in the interior of the surface has exactly six neighbors. The neighbors of vertex v_j are consistently ordered in a clockwise fashion. Vertices on the edge of a surface may have anywhere from two to five neighbors.

Neighboring vertices are further organized into triangular planar surface elements called *facets*, denoted $\mathbf{F} = (f_1, f_2, \ldots, f_{n_f})$. The vertices of a facet are also ordered in a clockwise fashion. In this work, we require that the initial estimate of the surface have facets whose sides are of equal length. The objective function described below tends to maintain this equality, but does not strictly enforce it. In the course of the optimization, we refine the mesh by iteratively subdiving the facets into four smaller ones whose sides are still of roughly equal length.

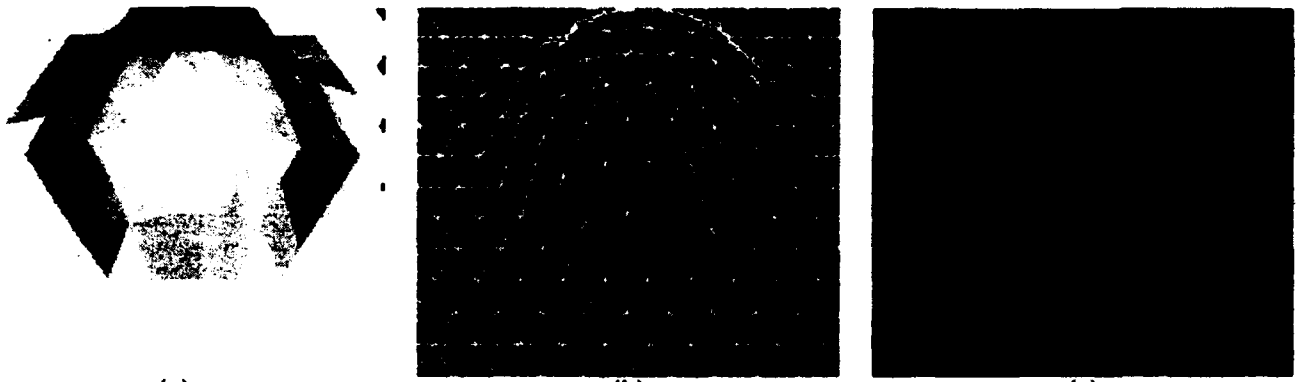


Figure 1: Projection of a mesh, and the Facet-ID image used to accommodate occlusions during surface reconstruction: (a) A shaded image of a mesh. (b) A wire-frame representation of the mesh (bold white lines) and the sample points in each facet (interior white points). (c) The Facet-ID image, wherein the color at a pixel is chosen to uniquely identify the visible facet at that point (shown here as a gray-level image).

In Figure 1, we show a shaded view and a wireframe representation of such a mesh. We also show what we call a "Facet-ID" image. For each input image, it is generated by encoding the index i of each facet f_i as a unique color, and projecting the surface into the image plane, using a standard hidden-surface algorithm. As discussed in Sections 2.3 and 2.4, we use it to determine which surface points are occluded in a given view and on which facets geometric constraints should be brought to bear.

2.1 Objective Function and Optimization Procedure

The objective function $\mathcal{E}(S)$ that we use to recover the surface is a sum of terms that take into account the image-based constraints—stereo and shape from shading—and the geometry-based constraints—features and silhouettes—that are brought to bear on the surface. To minimize $\mathcal{E}(S)$, we use an optimization method that is inspired by the heuristic technique known as a continuation method [Terzopoulos, 1986, Leclerc, 1989b, Leclerc, 1989a] in which we add a regularization term to the objective function and progressively reduce its influence. We define the total energy of the mesh, $\mathcal{E}_T(S)$, as

$$\begin{aligned} \mathcal{E}_T(S) &= \lambda_D \mathcal{E}_D(S) + \mathcal{E}(S) \\ \mathcal{E}(S) &= \sum_i \lambda_i \mathcal{E}_i(S) \ . \end{aligned} \tag{1}$$

The $\mathcal{E}_i(S)$ represent the image and geometry-based constraints, and the λ_i their relative weights. They are discussed in the following subsections. $\mathcal{E}_D(S)$, the regularization term, serves a dual purpose. First, we define it as a quadratic function of the vertex coordinates, so that it "convexifies" the energy landscape when λ_D is large and improves the convergence properties of the optimization procedure. Second, as shown previously [Fua and Leclerc, 1993], in the presence of noise, some amount of smoothing is required to prevent the mesh from overfitting the data, and wrinkling the surface excessively.

In our implementation, we take $\mathcal{E}_D$ to be a measure of the curvature or local deviation from a plane at every vertex. We approximate this as follows.

Consider a perfectly planar hexagonal mesh for which the distances between neighboring vertices are exactly equal. Recall that the mesh is defined so that the neighbors of a vertex v_i are ordered in a clockwise fashion, and are denoted $v_{N_i(j)}$. This notation is depicted in Figure 2(a). If the hexagonal mesh was perfectly planar, then the third neighbor over from the j^{th} neighbor, $v_{N_i(j+3)}$, would lie on a straight line with v_i and $v_{N_i(j)}$. Given that the inter-vertex distances are equal, this implies that coordinates of v_i equal the average of the coordinates of $v_{N_i(j)}$ and $v_{N_i(j+3)}$, for any j.

Given the above, we can write a measure of the deviation of the mesh from a plane as follows:

$$\mathcal{E}_D(S) = \sum_{i=1}^{n_v} \sum_{\substack{j=1 \\ k=N_i(j) \\ k'=N_i(j+3)}}^{3} (2x_i - x_k - x_{k'})^2 + (2y_i - y_k - y_{k'})^2 + (2z_i - z_k - z_{k'})^2 \tag{2}$$

Note that this term is also equivalent to the squared directional curvature of the surface when the sides have equal lengths [Kass *et al.*, 1988]. This term can be made to accommodate multiple resolutions of facets by normalizing each term by the nominal intervertex spacing of the facets.

In previous implementations [Fua and Leclerc, 1993], we have performed optimization using a standard conjugate-gradient descent procedure [Press *et al.*, 1986]. However, the $\mathcal{E}_D$ term described here is amenable to a "snake-like" optimization technique [Kass *et al.*, 1988]. We embed the curve in a viscous medium and solve the equation of dynamics

$$\begin{aligned} \frac{\partial \mathcal{E}_T}{\partial S} + \alpha \frac{dS}{dt} &= 0, \\ \text{with } \frac{\partial \mathcal{E}_T}{\partial S} &= \frac{\partial \mathcal{E}_D}{\partial S} + \frac{\partial \mathcal{E}}{\partial S}, \end{aligned} \tag{3}$$

where $\mathcal{E}_T$ is the total energy of Equation 1, α the viscosity of the medium, and S the state vector that defines the current position of the mesh that is the vector of the x,y, and z coordinates of the vertices. Since the

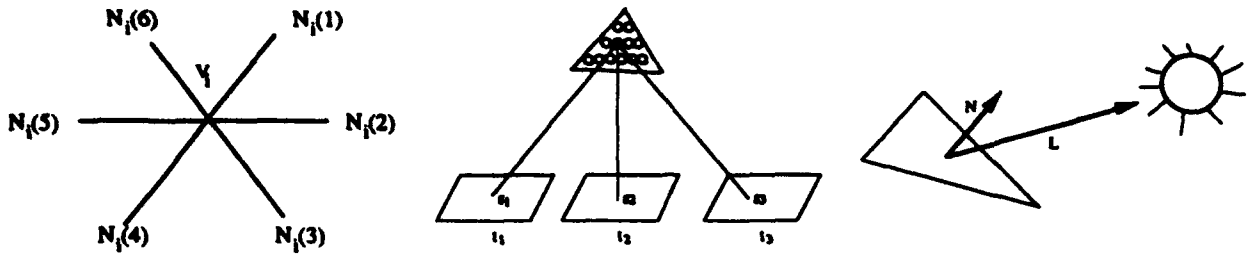


Figure 2: Vertices and facets of a mesh: (a) The six neighbors $N_i(j)$ of a vertex v_i are ordered clockwise. (b) Facets are sampled at regular intervals as illustrated here. The stereo component of the objective function is computed by summing the variance of the gray level of the projections of these sample points, the g_is. (c) The albedo of each facet is estimated using the facet normal $\vec{N}$, the light source direction $\vec{L}$, and the average gray level of the projection of the facet into the images. The shading component of the objective function is the sum of the squared differences in estimated albedo across neighboring facets.

deformation energy $\mathcal{E}_D$ in Equation 2 is quadratic, its derivative with respect to S is linear, and therefore Equation 3 can be rewritten as

$$\begin{aligned} K_S S_t + \alpha(S_t - S_{t-1}) &= -\left.\frac{\partial \mathcal{E}}{\partial S}\right|_{S_{t-1}} \\ \Rightarrow (K_S + \alpha I) S_t &= \alpha S_{t-1} - \left.\frac{\partial \mathcal{E}}{\partial S}\right|_{S_{t-1}} \end{aligned} \tag{4}$$

where

$$\frac{\partial \mathcal{E}_D}{\partial S} = K_S S,$$

and K_S is a sparse matrix. Note that the derivatives of $\mathcal{E}_D$ with respect to x,y, and z are decoupled so that we can rewrite Equation 4 as a set of three differential equations in the three spatial coordinates:

$$\begin{aligned} (K + \alpha I) X_t &= \alpha X_{t-1} - \left.\frac{\partial \mathcal{E}}{\partial X}\right|_{X_{t-1}} \\ (K + \alpha I) Y_t &= \alpha Y_{t-1} - \left.\frac{\partial \mathcal{E}}{\partial Y}\right|_{Y_{t-1}} \\ (K + \alpha I) Z_t &= \alpha Z_{t-1} - \left.\frac{\partial \mathcal{E}}{\partial Z}\right|_{Z_{t-1}} \end{aligned}$$

where X,Y, and Z are the vectors of the x,y and z coordinates of the vertices, and K is a sparse matrix. In fact, for our hexagonal meshes, K turns out to be a banded matrix and this set of equations can be computed efficiently using LU decomposition and backsubstitution. Note that the LU decomposition need be recomputed only when α changes. When α is constant, only the backsubstitution step is required. In practice α is computed automatically at the start of the optimization procedure so that a prespecified average

vertex motion amplitude is achieved [Fua and Leclerc, 1990]. The optimization proceeds as long as the total energy decreases; when it increases the algorithm backtracks and increases α, thereby decreasing the step size.

In general we optimize all three spatial components simultaneously. However, when dealing with surfaces for which motion in one direction leads to more dramatic changes than motions in others, as is typically the case with the z direction in Digital Elevation Models (DEMs), we have found the following heuristic to be useful. We first fix the x and y coordinates of vertices and adjust z alone. Once the surface has been optimized, we then allow all three coordinates to vary.

2.2 Combining the Components

The total energy of Equation 1 is a sum of terms whose magnitudes are image- or geometry-dependent and are therefore not necessarily commensurate. One therefore needs to scale them appropriately, that is to define the λ weights so as to make the magnitude of their contributions commensurate and independent of the specific radiometry or geometry of the scene under consideration.

From Equation 4, it can be seen that the dynamics of the optimization are controlled by the gradient of the objective function. As a result, we have found that an effective way to normalize the contributions of the various components of the objective function is to define a set of user-specified weights λ_i' such that

$$\sum_{1 \leq i \leq n} \lambda_i' < 1 \, .$$

These weights are then used to define the λs as follows

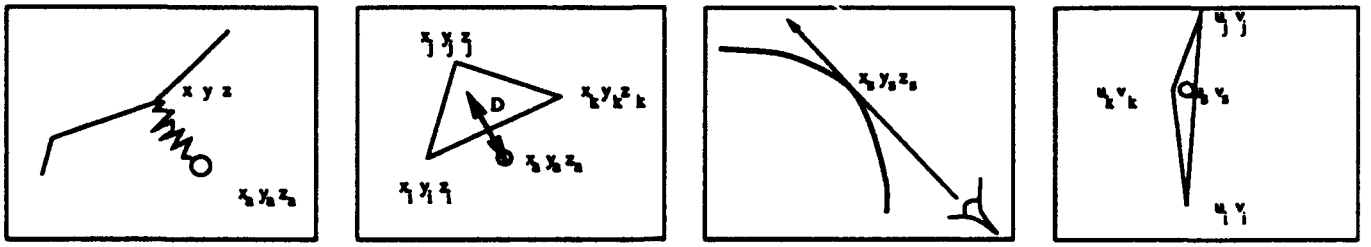
$$\begin{aligned} \lambda_i &= \frac{\lambda_i'}{\| \vec{\nabla} \mathcal{E}_i(\mathcal{S}^0) \|} \\ \lambda_D &= \frac{\lambda_D'}{\| \vec{\nabla} \mathcal{E}_D(\mathcal{S}^0) \|} \end{aligned} \tag{5}$$

where $\lambda_D' = 1 - \sum_i \lambda_i'$ and $\mathcal{S}^0$ is the surface estimate at the start of each optimization step. We first proposed this normalization scheme in [Fua and Leclerc, 1990], and it is analogous to standard constrained optimization techniques in which the various constraints are scaled so that their eigenvalues have comparable magnitudes [Luenberger, 1984]. In practice we have found that, because the normalization makes the influence of the various terms comparable irrespective of actual radiometry or dimensions, the user-specified λ_i' weights are context-specific but not image-specific. In other words, we use one set of parameters for images of faces when combining stereo, shape-from-shading, and silhouettes, and another when dealing with aerial images of terrain using stereo and 3-D point constraints, but we do not have to change them for different faces or different landscapes.

The continuation method discussed in Section 2.1 is implemented by taking the initial value of λ_D' to be 0.5 and then progressively decreasing it while keeping the relative values of the λ_i's constant.

2.3 *Geometric Constraints*

We have explored the constraints generated by 3–D points, 3–D linear features, and 2–D silhouettes. 3-D points are treated as attractors that are connected to the surface by springs, and 3–D linear features are taken to be collections of such points. They are described in Section 2.3.1. 2–D silhouettes are image features that constrain the surface normals and are discussed in Section 2.3.2.



(a) (b) (c) (d)

Figure 3: 3–D and 2–D point constraints: (a) Point attractor modeled as a spring attached to a vertex. (b) Point attractor modeled as a spring attached to the closest surface point. (c) Occlusion contours are the locus of the projections of the (x_s, y_s, z_s) surface points for which a camera ray is tangential to the surface. (d) In practice, the (u_s, v_s) projection of such a point must be colinear with the projections of the vertices of the facet that produces the observed silhouette edge.

2.3.1 3–D Points

3-D attractors can be handled in two different ways. The easiest way is to add the following term to the objective function for each attractor (x_a, y_a, z_a):

$$e_a = 1/2((x_a - x_i)^2 + (y_a - y_i)^2 + (z_a - z_i)^2) \tag{6}$$

where x_i,y_i, and z_i are the coordinates of the closest mesh vertex, as shown in Figure 3(a). This term models each attractor as a spring. These simple springs could be replaced by more sophisticated "breakable" springs such as those used by Terzopoulos *et al.* [1991] to account for possible outliers. This formulation has proved useful to interpolate dense and relatively error-free range maps produced using either correlation-based stereo or laser range maps: we project the vertices of the mesh into the map and compute x_a, y_a, z_a wherever range data are available.

The above is, however, inadequate if one wishes to use facets that are large enough so that attracting the vertices, as opposed to the surface point closest to the attractor, would cause unwarranted deformations of the mesh. This is especially important when using a sparse set of attractors. In this case, the energy term of Equation 6 must be replaced by one that attracts the surface without warping it. In our implementation, this is achieved by redefining e_a as

$$e_a = 1/2d_a^2 \tag{7}$$

where d_a is the orthogonal distance of the attractor to the closest facet. The normal vector to a facet can be computed as the normalized cross product of the vectors defined by two sides of that facet, and d_a as the dot product of this normal vector with the vector defined by one of the vertices and the attractor. Letting $(x_i, y_i, z_i)_{1 \leq i \leq 3}$ be the three vertices of a facet, consider the polynomial D defined as

$$D = \begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_a & y_a & z_a & 1 \end{vmatrix}$$

$$= C_x x + C_y y + C_z z$$

where C_x,C_y, and C_z are polynomial functions of x_i,y_i, and z_i. It is easy to show that the facet normal is parallel to the vector (C_x, C_y, C_z) and that the square of the orthogonal distance d_s^2 of the attractor to the facet can be computed as

$$d^2 = D^2/(C_x^2 + C_y^2 + C_z^2)$$

Finding the "closest facet" to an attractor is computationaly expensive in general. However, in our specific case the search can be made efficient and fast if we assume that the 3–D points can be identified by their projection in an image. We project the mesh in that image, generate the corresponding Facet-ID image—which must be done in any case for other computations—and look up the facet number of the point's projection. This applies, for example, to range maps, edge- or correlation-based stereo data, and hand-entered features that can be overlaid on various images. We typically recompute the facet attachments at every iteration of the optimization procedure so as to allow facets to slide as necessary. Since the points can potentially come from any number of such images, this method can be used to fuse 3–D data from different sources.

2.3.2 Silhouettes

Contrary to 3–D edges, silhouette edges are typically 2–D features since they depend on the viewpoint and cannot be matched across images. However, as shown in Figure 3(c), they constrain the surface tangent. Each point of the silhouette edge defines a line that goes through the optical center of the camera and is tangent to the surface at its point of contact with the surface. The points of a silhouette edge therefore define a ruled surface that is tangent to the surface. In terms of our facetized representation, this can be expressed as follows. Given a silhouette point (u_s, v_s) in an image, there must be a facet with vertices $(x_i, y_i, z_i)_{1 \leq i \leq 3}$ whose image projections $(u_i, v_i)_{1 \leq i \leq 3}$, as well as (u_s, v_s), all lie on a single line as depicted by Figure 3(d). This implies that the three determinants of the form

$$\begin{vmatrix} u_i & u_j & u_s \\ v_i & v_j & v_s \\ 1 & 1 & 1 \end{vmatrix}, \quad 1 \leq i \leq 3, i < j \leq 3$$

must be equal to zero. We enforce this for each silhouette point by adding to the objective function a term of the form

$$e_s = 1/2 \sum_{1 \leq i \leq 3, i < j \leq 3} \begin{vmatrix} u_i & u_j & u_s \\ v_i & v_j & v_s \\ 1 & 1 & 1 \end{vmatrix}^2 \tag{8}$$

where the (u_i, v_i)s are derived from the (x_i, y_i, z_i) using the camera model.

As with the 3–D attractors described in Section 2.3.1, the main problem is to find the "silhouette facet" to which the constraint applies. Since the silhouette point (u_s, v_s) can lie outside the projection of the current estimate of the surface, we search the Facet-ID image in a direction normal to the silhouette edge for a facet that minimizes e_s and that is therefore the most likely to produce the silhouette edge. This, in conjunction with our coarse-to-fine optimization scheme, has proved a robust way of determining which facets correspond to silhouette points.

2.4 Image Constraints

In this work, we use two complementary image-based constraints: stereo and shape-from-shading.

The stereo component of the objective function is derived by comparing the gray-levels of the points in all of the images for which the projection of a given point on the surface is visible, as determined using the

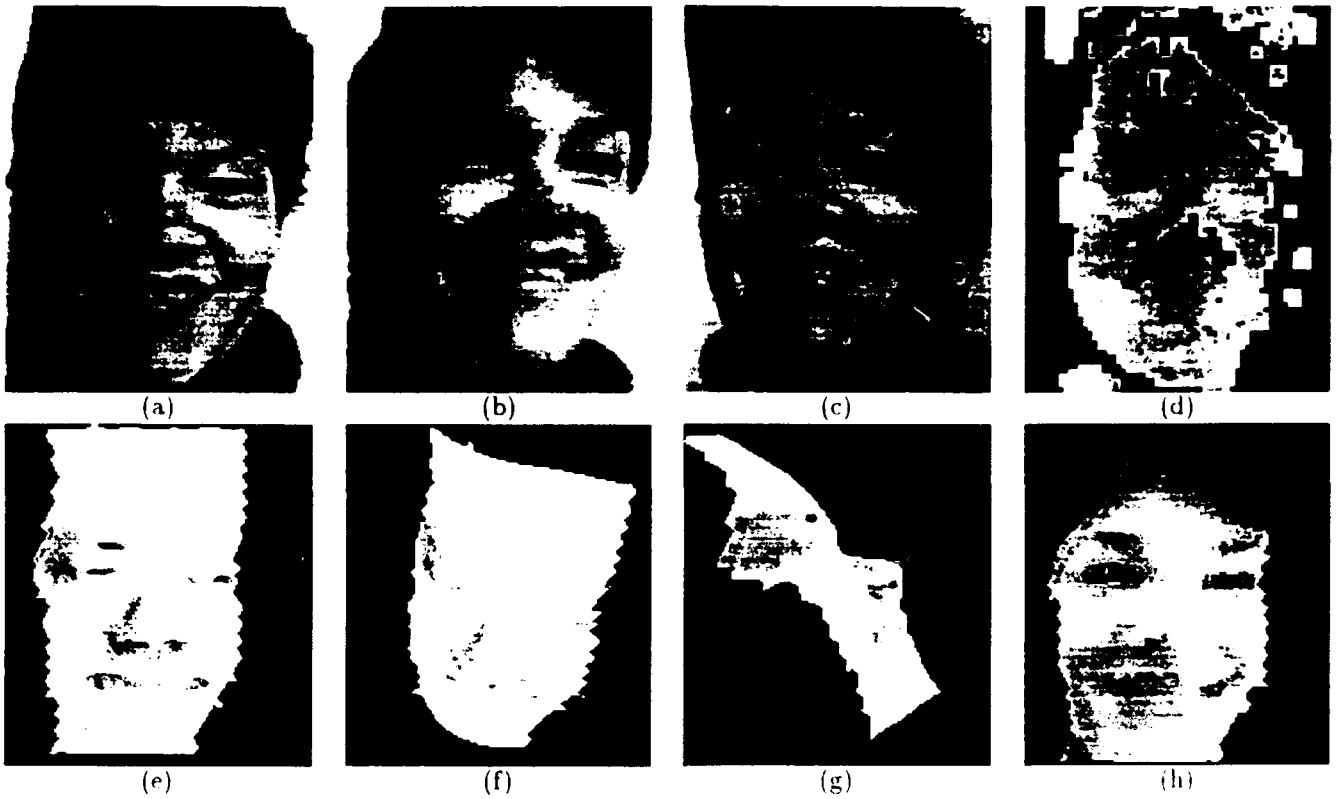


Figure 4: Recovering the shape of a face by combining stereo and shape-from-shading. (a) (b) (c) Triplet of face images (courtesy of INRIA). (d) Disparity map. (e) (f) (g) Shaded views of the reconstructed surface after optimization. (h) The recovered albedo map.

Facet-ID image. This comparison is done for a uniform sampling of the surface, as shown in Figure 2(b). This method allows us to deal with arbitrarily slanted regions and to discount occluded areas of the surface.

The shading component, depicted in Figure 2(c), of the objective function is computed using a method that does not invoke the traditional assumption of constant albedo. Instead, it attempts to minimize the variation in albedo across the surface, and can therefore deal with both constant albedo surfaces as well as surfaces whose albedo varies slowly.

Stereo information is very robust in textured regions but potentially unreliable elsewhere. We therefore use it mainly in textured areas by weighting the stereo component most strongly for facets of the triangulation that project into textured image areas. Conversely, the shading information is more reliable where there is little texture and is weighted accordingly.

These two terms are central to our approach: they are the ones that allow the combination of geometric information with image information. However, since their behavior and implementation have already been extensively discussed elsewhere, we do not describe them any further here and refer the interested reader to our previous publication [Fua and Leclerc, 1993]. In Figure 4, we show the reconstruction of a face using only stereo and shape-from-shading.

3 Applications

Our framework allows us to combine geometric constraints with image-based constraints to derive surface reconstructions and to refine previously computed surfaces. In this section, we demonstrate its capabilities using difficult imagery.

3.1 From 3–D Constraints to Detailed Surfaces

Our system deals with the various sources of 3–D information, whether dense, such as range maps or correlation-based stereo disparity maps, or linear, such as hand-entered features or edge-based stereo disparity maps, in the same fashion. Both are sampled at regular intervals to generate collections of 3–D attractors that are used to define energy terms using Equation 6 or 7.

Especially in the case of sparse features, the "snake-type" optimization technique of Section 2.1 has proved more effective than more classical techniques such as conjugate gradient at propagating constraints across the mesh. To further speed the computation and prevent the mesh from becoming stuck in undesirable local minima, we use several levels of mesh size—three in the examples shown below—to perform the computation. We start with a relatively coarse mesh that we optimize. We then refine it by splitting every facet into four smaller ones and reoptimizing. Finally, we repeat the split and optimization processes one more time.

3.1.1 Dense 3–D Data

In Figure 5, we show an image of a face and a corresponding range map computed using structured light. Although it is fairly accurate, this particular method introduces artifacts that are highlighted in Figure 5(c). We first fit a surface to these points by starting from a flat surface and taking the total energy $\mathcal{E}_T$ of Equation 1 to be

$$\begin{aligned} \mathcal{E}_T &= \lambda_D \mathcal{E}_D + \lambda_A \mathcal{E}_A \\ &= \lambda_D \mathcal{E}_D + \lambda_A \sum_a e_a \end{aligned} \tag{9}$$

where the e_a are defined for each range-data point as the attraction terms of Equation 7. Because of the artifacts of the original range data, the resulting surface is approximately correct but excessively wrinkly, as shown in Figure 5(d) and (e). Of course, we could simply smooth the surface but we would then be at risk of losing important details such as the mouth or the fine structures on the side of the nose. Our approach provides us with a better way of dealing with this problem: we can fuse the range information with the shading information of the intensity image of Figure 5(a). To do so, we add to $\mathcal{E}_T$ the shading term defined in Section 2.4, that we denote $\mathcal{E}_{Sh}$:

$$\mathcal{E}_T = \lambda_D \mathcal{E}_D + \lambda_A \mathcal{E}_A + \lambda_{Sh} \mathcal{E}_{Sh}.$$

We restart the optimization from the flat initial surface. The new surface, shown in Figure 6, is much smoother, but the mouth is well preserved and the side of the nose better defined. Note, however, that in the side views the bottom of the nose is not flat enough. This is not surprising since the shading information is of no use there. We address this problem in Section 3.2.

3.1.2 Sparse 3–D Data

We now turn to sparse 3–D data. In Figure 7, we show a stereo pair of a rock outcrop forming an almost vertical cliff. Note that, even though the geometry is almost epipolar, these two images are very hard to fuse both for humans and for automated procedures. In Figure 7(c), we show the output of a correlation result

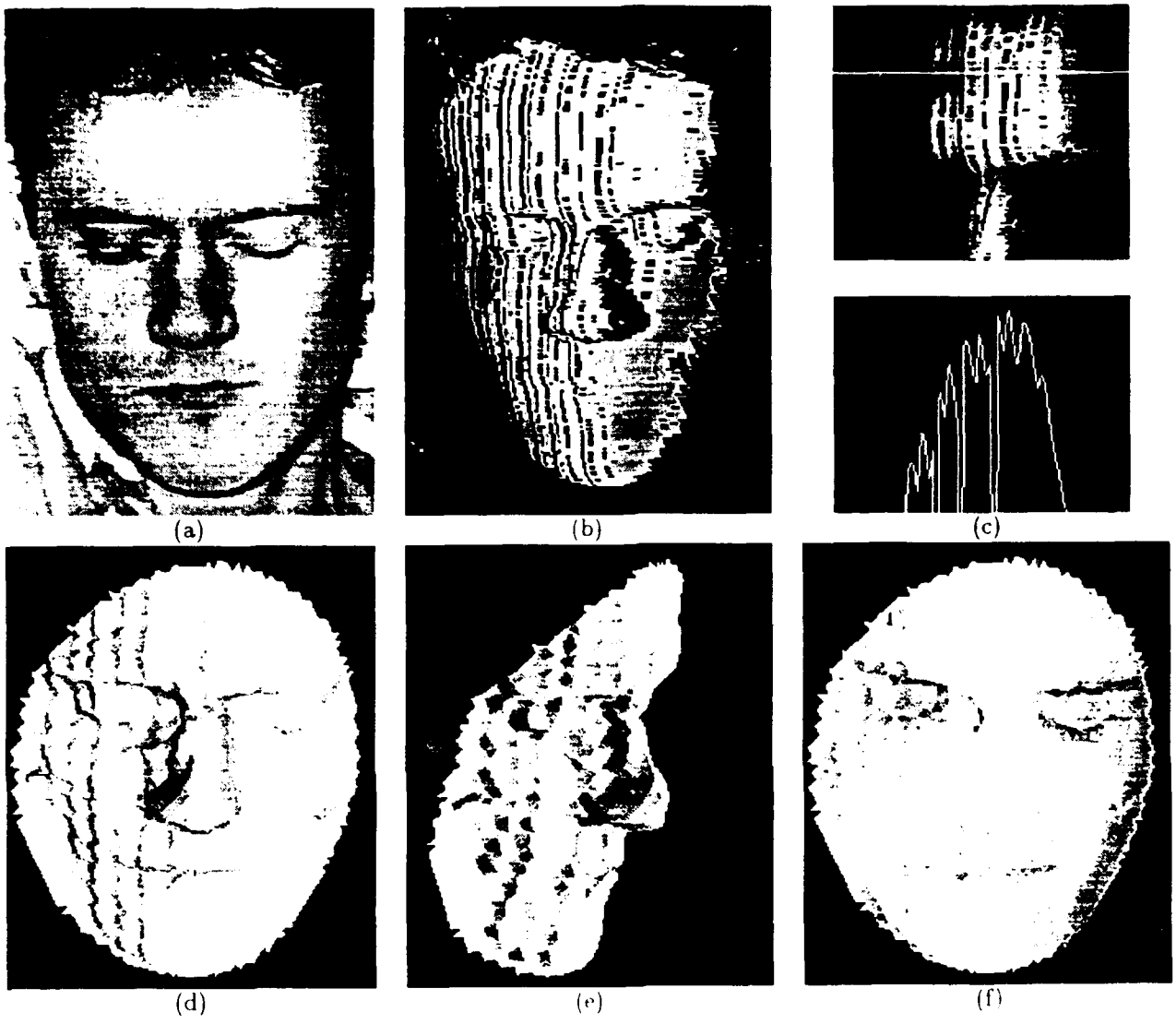


Figure 5: Fitting a surface to range data: (a) Image of a face (Courtesy of ETH Zurich) (b) Corresponding range image computed using structured light. (c) A window of the range image in which gray levels have been stretched to emphasize the vertical wrinkles and the histogram of a horizontal slice. (d) (e) Shaded views of the surface reconstructed by using the range-data points as attractors. (f) The corresponding albedo map.

[Fua, 1993] that gives no information about the shape of the outcrop. This can be attributed to the fact that, in the cliff area, the fundamental assumption underlying correlation-based stereo using a fixed-shape window is violated: the depth is not constant within a correlation window. To demonstrate the data-fusion capabilities of our approach, we supply the 3-D edges whose projections are shown in Figure 7(d) and (e). To do so, we have used the 3-D snakes [Fua and Leclerc, 1990] that are embedded in the SRI Cartographic

Figure 6: Combining range-data with shape-from-shading information: (a)(b)(c) Shaded views of the refined reconstruction of the face of Figure 5 using shading.

Modeling Environment (CME) [Quam and Strat, 1991]: rough contours are hand-entered and treated as the projections of polygonal 3-D curves whose x,y, and z coordinates are then optimized to maximize the average edge-strength along the projections. Alternatively, we could take advantage of the output of 3-D edge detectors such as those described in [Ayache and Lustman, 1987, Robert and Faugeras, 1991, Ma and Thonnat, 1992, Meygret *et al.*, 1990].

By using the energy term of Equation 9, we attract an initially flat surface to both the stereo data and the 3-D outlines and produce a shape estimate that is roughly correct but much too smooth, as can be seen in Figure 8 (b) and (c).

By adding either the stereo term alone to $\mathcal{E}_T$, Figure 8 (d), or both the stereo and shading terms, Figures 8 (e) and (f), we can generate a much more realistic model of the surface. Note that in Figure 8 (e) the cracks in the right side of the outcrop are well modeled. Our object-centered representation has no trouble accomodating the sharply slanted surfaces.

In Figure 9, we show another application of our technique in a semiurban environment using images of a model board. We have used the 3-D snakes to outline some of the linear features visible in the images. We then generate the rough estimate of the surface shape of Figure 10(b), and improve it using stereo as shown in Figure 10(c). In addition, we have used CME to model the buildings as extruded objects. We exploit them to mask out occluded areas when computing the stereo energy. This is achieved naturally in our system by using the projections of the building models in each view to zero out the corresponding Facet-ID image. In this way, the facet samples that project at these locations are discounted during the computation of the stereo energy defined in Section 2.4. Since buildings cannot be very well described by our smooth mesh, ignoring those pixels amounts to assuming that the terrain is smooth below the buildings and prevents the surface from wrinkling unduly.

3.2 Refining Previously Derived Models

So far, we have shown how our technique can be used to generate surface models "from scratch." However, very few vision algorithms—ours being no exception—consistently provide a perfect answer across scenes

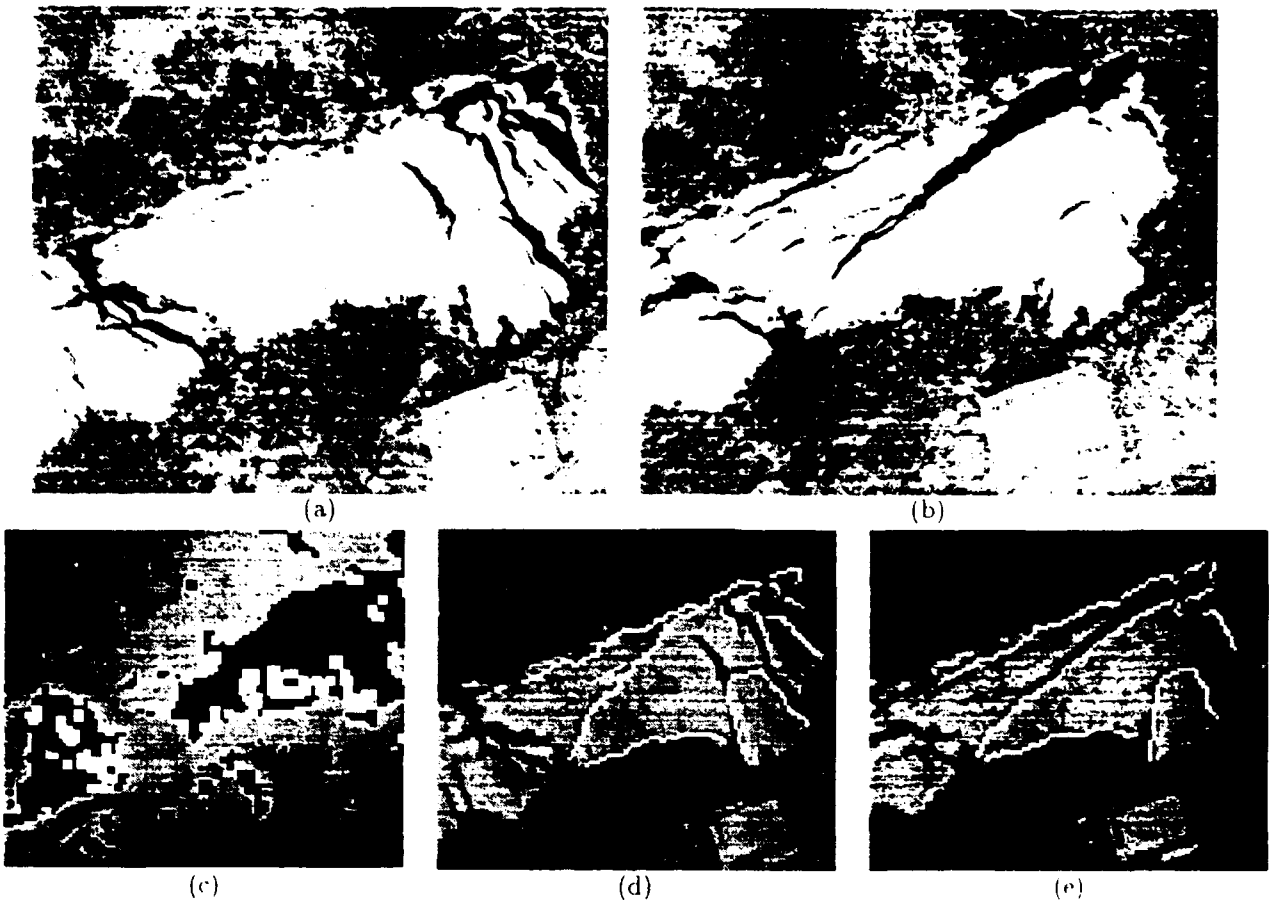


Figure 7: Semiautomated cartography of a rugged site: (a) (b) A hard-to-fuse stereo pair of a rock outcrop with an almost vertical cliff. (c) Disparity map. Within the outcrop the correlation-based algorithm provides almost no information; outside of it the terrain is almost flat. (d) (e) The projections of a few 3-D features outlined using 3-D snakes.

using a predetermined set of information sources and analysis parameters. For applications such as cartography or 3-D graphics, it is often important to be able to easily refine a previously derived result, such as an old DEM or the output of a fully automated procedure, using additional clues. This can be done using both 3-D contours and silhouettes.

We start with an example involving the two aerial images of Figure 11, at the top of which is a very sharp cliff that casts shadows on the ground. Starting from a coarse and inaccurate DEM, we generate the surface shown in Figure 11(e), using stereo alone. By computing the disparities associated with that improved model, we have visually checked that it is correct except in the immediate vicinity of the cliff, where it is too smooth. This should be expected: our objective function $\mathcal{E}_I$ includes a smoothness term, and the face of the cliff is not visible in those images and therefore provides no stereo clues. By sketching the edge of the cliff and the shadows with our 3-D snakes and using them to add an attraction term to the objective function, we can deform the surface slightly to produce the result shown in Figure 11(f) where the

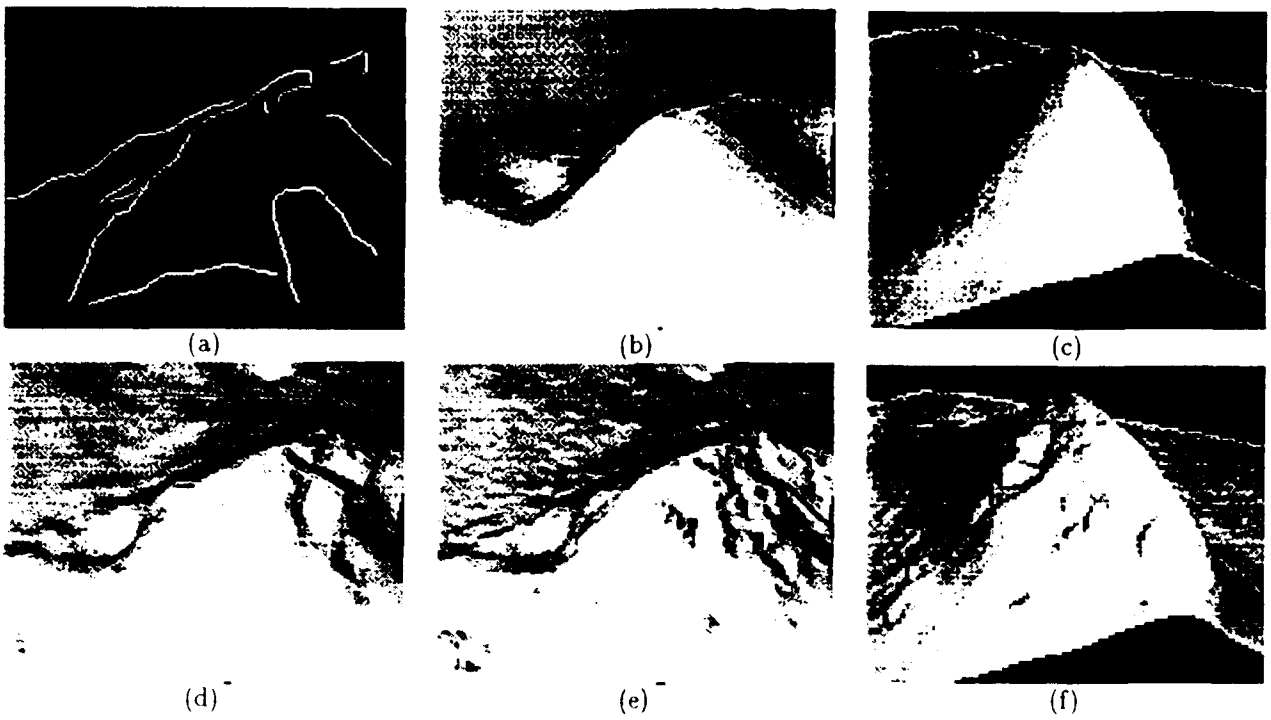


Figure 8: Combining 3-D constraints with stereo and shape-from-shading: (a) The recovery of the terrain for the aerial scene of Figure 7 starts with a flat surface that is attracted by the 3-D outlines and the 3-D cloud of points corresponding to the disparity map. (b) (c) Shaded views of the reconstructed surface using only those constraints. (d) Refinement using stereo (e) (f) Refinement using both stereo and shape from shading.

ridge is better defined. To further check the validity of our result, we have used the known sun direction to predict which parts of the ground are in shadow. To do this we generate a sun-view, that is an orthographic view as seen from the sun's viewpoint, and the corresponding Facet-ID image. For every facet, we compute the proportion of samples that are visible in this sun view as shown in Figure 11(g). The facets for which a large proportion of samples is occluded are those in shadow. As can be seen, these shadowed facets match the actual shadows fairly well, which leads us to believe that our reconstruction is accurate.

Silhouettes are also very good indicators of the quality of a reconstruction. For example the reconstruction of the bottom of the nose in Figure 6 is not quite right as evidenced by its silhouette in the side view of the same man shown in Figure 12. However, we can use the silhouette constraints of Section 2.3.2 with the two silhouettes shown in the figure. The silhouettes are 2-D curves that have been outlined using 2-D snakes. In the manner of Section 3.1.1, we take the total energy $\mathcal{E}_T$ to be

$$\begin{aligned} \mathcal{E}_T &= \lambda_D \mathcal{E}_D + \lambda_S \mathcal{E}_S + \lambda_{Sh} \mathcal{E}_{Sh} \\ \mathcal{E}_S &= \sum_s e_s \end{aligned}$$

where the e_s are the silhouette attraction terms of Equation 7 and $\mathcal{E}_{Sh}$ the shading term described in

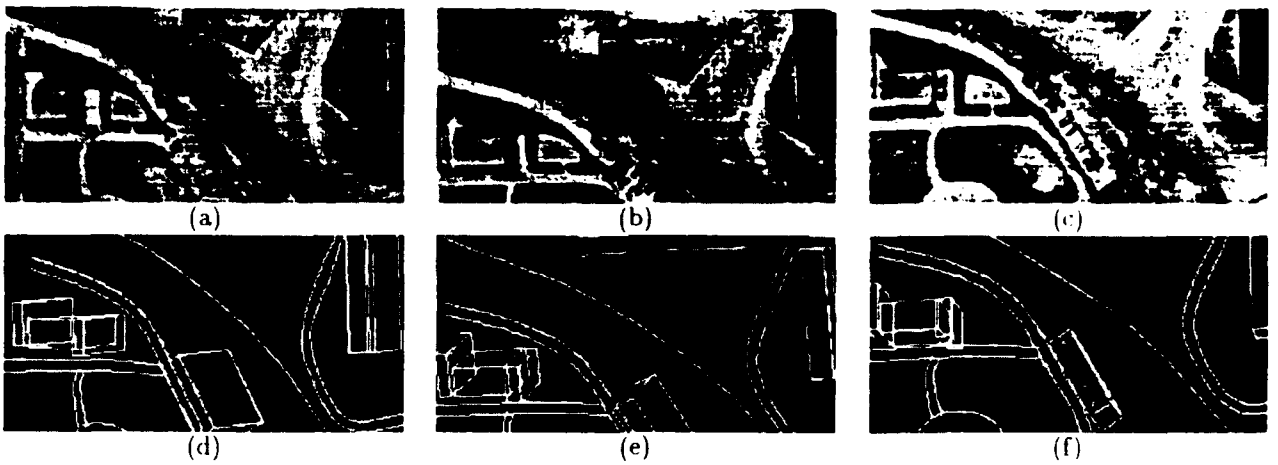


Figure 9: Semiautomated cartography of a semiurban site: (a) (b) (c) Three images taken with different light source directions. (d) (e) (f) Projections of hand-entered 3-D linear features and building blocks. Note that the bases of the buildings extend below the ground

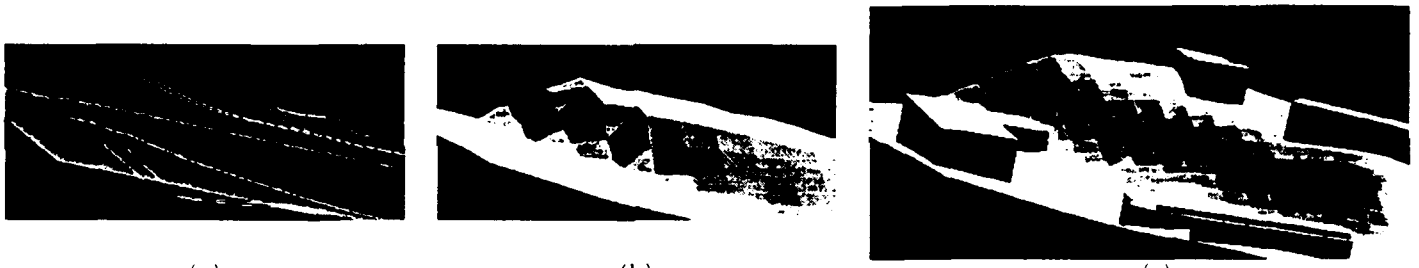


Figure 10: Combining 3-D constraints and visibility constraints with stereo. (a) The 3-D linear features of Figure 9 above the flat plane used as the initial surface estimate. (b) A rough estimate of the ground-level surface (c) Surface after optimization using both stereo and hand-entered buildings to mask occluded areas.

Section 2.4. We use these terms to deform the nose region and generate the improved result shown in Figure 12(c).

The face reconstruction of Figure 4 presents us with a slightly different problem. We have used a correlation-based stereo algorithm to provide us with an initial estimate. This algorithm gave us no information on the sharply slanted parts of the face, which are therefore missing from the reconstruction. The silhouettes of the face, however, are clearly visible in Figure 13 and easy to outline. To take advantage of these, we again use a coarse to fine strategy. We start with a larger and coarser mesh that evolves under the influence of the silhouettes and the vertices of the original reconstruction that are treated as attractors. When the mesh has been refined and optimized, we complete the optimization procedure by turning on the

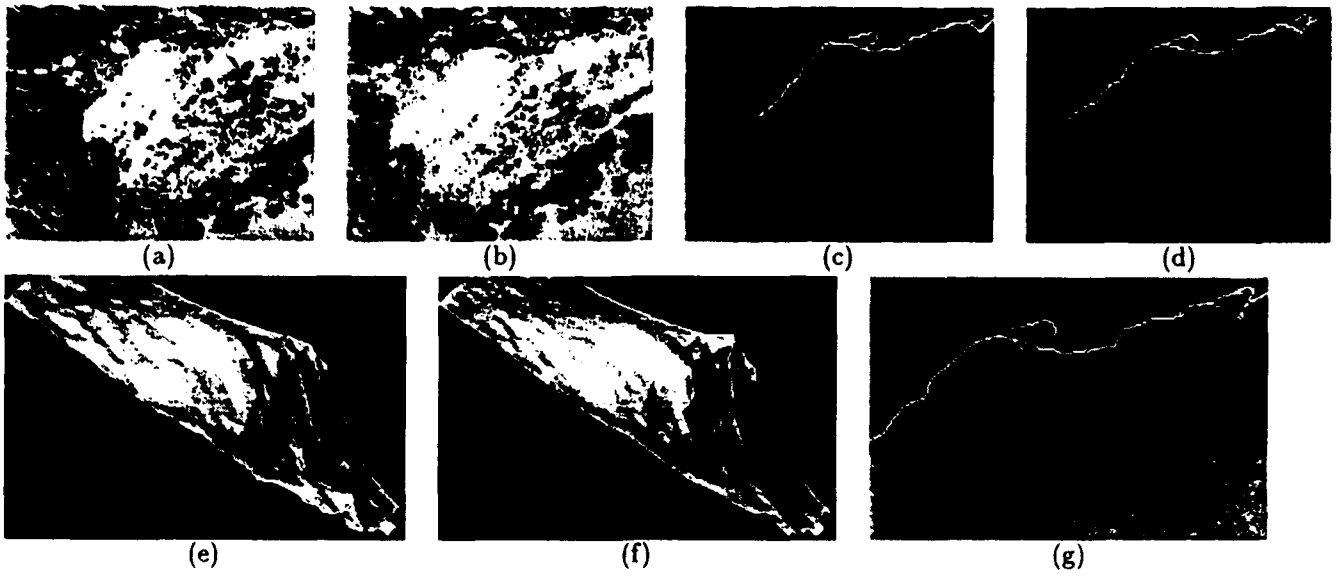


Figure 11: Improving and checking a DEM: (a) (b) An aerial stereo pair of a cliff with clearly visible shadows. (c) (d) The cliff's ridge and cast shadows outlined using 3–D snakes. (e) Reconstructed surface using stereo alone. (f) Reconstructed surface using both stereo and the 3–D outlines as attractors. (g) Predicted shadow areas in black. The prediction was carried out using the reconstruction shown in (f) and the known sun direction. Note that these hypothesized shadows closely match the actual ones. Note also that, were we to use the original reconstruction shown in (e) to perform this computation, no shadows would be predicted because the surface is too smooth.

full objective function:

$$\mathcal{E}_T = \lambda_D \mathcal{E}_D + \lambda_S \mathcal{E}_S + \lambda_{Sh} \mathcal{E}_{Sh} + \lambda_{St} \mathcal{E}_{St},$$

where $\mathcal{E}_{Sh}$ and $\mathcal{E}_{St}$ denote the shading and stereo terms presented in Section 2.4. The results are shown in Figure 13(c),(d) and (e).

The silhouettes used in the two examples above have been entered semi-automatically. But here again, we could take advantage of automatically extracted ones [Cipolla and Blake, 1990, Liedtke *et al.*, 1991, Vaillant and Faugeras, 1992].

4 Conclusion

We have presented a surface reconstruction method that uses an object-centered representation to recover 3–D surfaces. Our method uses both monocular shading cues and stereoscopic cues from any number of images while correctly handling self-occlusions. It can also take advantage of the geometric constraints derived from measured 3–D points and 2–D silhouettes. These complementary sources of information are combined in a unified manner so that new ones can be added easily as they become available.

Using a variety of real imagery, we have demonstrated that the resulting method is quite powerful and flexible, allowing for both completely automatic reconstruction in straightforward circumstances, and for

(a) (b) (c)

Figure 12: Using silhouettes to improve a reconstruction: (a) The face of Figure 5 with a silhouette at the bottom of the nose outlined. (b) A side view of the same face with a second nose silhouette. (c) Shaded views of the refined reconstruction using both shading and the two silhouettes.

user-assisted reconstruction in more complex circumstances. User assistance is provided primarily through the introduction and identification of a small number of hand-entered linear and point features using semi-automated "snake" technology. The method is also controlled by a small number of parameters that specify the relative importance of the various information sources. These parameters typically do not need to be adjusted for images within a given class (such as images of faces or high-altitude aerial images), but only across classes.

The method has valuable capabilities for applications such as 3-D graphics model generation and high-resolution cartography in which a human can select the sources of information to be used and their relative importance. For example, in the case of mapping, one must ensure that the terrain model conforms to the feature data and does not violate any physical constraints: roads should be on the ground and not overly tilted, streams should stay within stream beds, buildings should not be floating in space, and so on. Our method allows one to both satisfy these constraints and account as well as possible for the observed image data.

In future work, we will study in a more quantitative manner the influence of the various terms of our objective function and their relative weights. This will require the use of ground-truth and carefully controlled conditions. We plan to set up a facility that will allow us to acquire the necessary data. We will also strive to replace some of the hand-entered geometric cues by automatically extracted ones and to investigate more complex topologies than the ones shown here. A principled way to do so would be to rephrase our modeling task as one of finding the "best" description of a scene in terms of the Minimum Length Description (MDL) principle [Rissanen, 1987, Leclerc, 1989b, Fua and Hanson, 1991]. It can be shown that the objective function that we propose here can be reformulated in terms of the MDL principle. After optimization using stereo and shape from shading, the surface ought to provide the best possible compromise between simplicity of

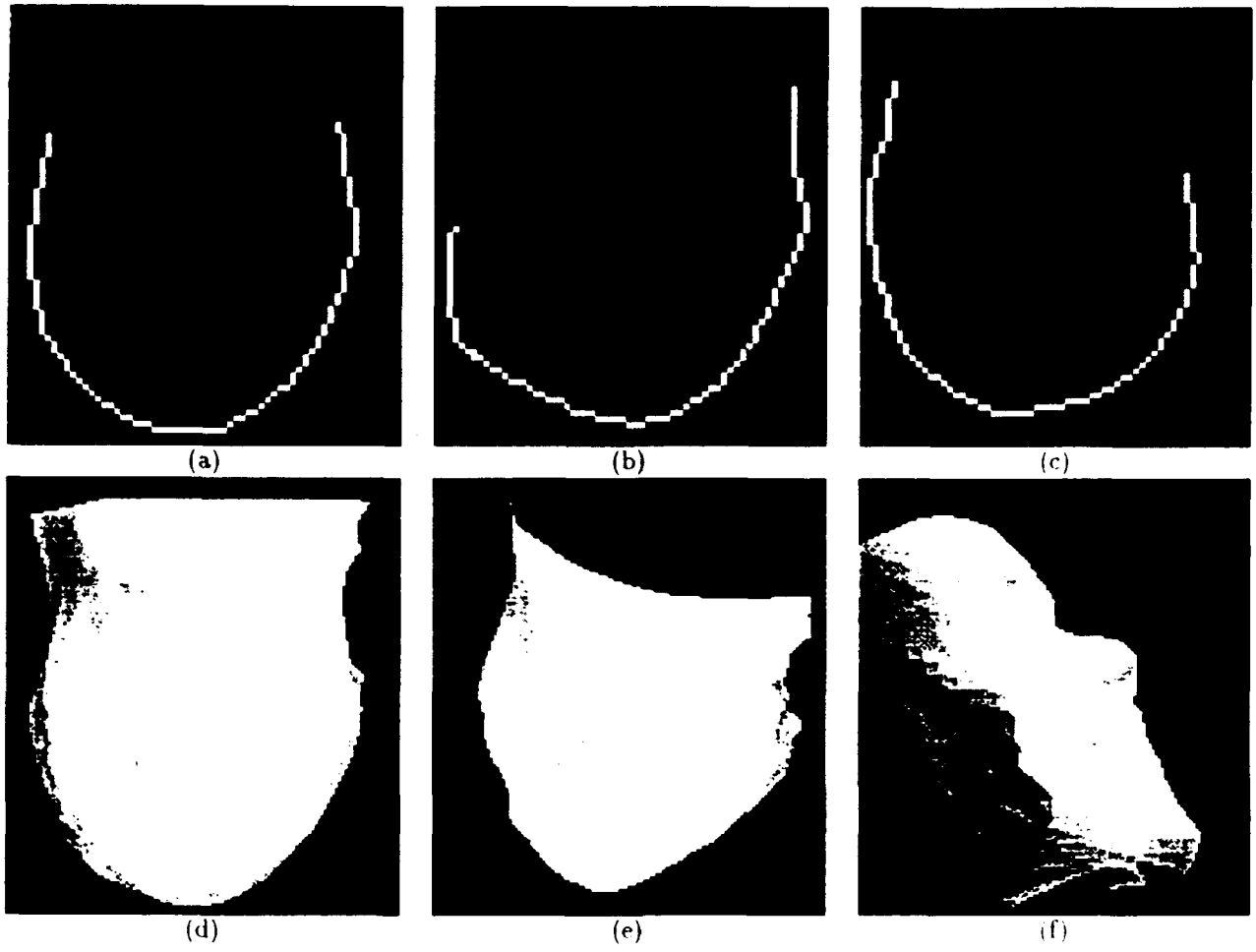


Figure 13: Using silhouettes to expand the scope of our method: (a) (b) (c) Silhouettes of the face in the three views of Figure 4 outlined using 2-D snakes. (d) (e) (f) Shaded views of reconstructed surface after optimization using stereo, shading, and the constraints provided by the silhouettes.

description of the surface and fit to the image data in terms of the simple vocabulary of triangulated meshes. The extensions that we have described above allow us to enrich the vocabulary by adding new primitives – ridges, building, roads, and so on –that allow an even more effective description. This approach would give us a principled way to accept or reject new objects in our overall representation.

Acknowledgments

Support for this research was provided by various contracts from the Advanced Research Projects Agency. We wish to thank Hervé Matthieu, Olivier Monga, Olaf Kubler and Marjan Trobina from INRIA and ETH

Zurich who have provided us with the face images and corresponding calibration data that appear in this paper and have proved extremely valuable to our research effort.

References

[Ayache and Lustman, 1987] N. Ayache and F. Lustman. Fast and reliable passive trinocular stereovision. In *International Conference on Computer Vision*, June 1987.

[Blake *et al.*, 1985] A. Blake, A. Zisserman, and G. Knowles. Surface descriptions from stereo and shading. *Image Vision Computation*, 3(4):183–191, 1985.

[Cipolla and Blake, 1990] R. Cipolla and A. Blake. The dynamic analysis of apparent contours. In *International Conference on Computer Vision*, 1990.

[Cohen *et al.*, 1991] I. Cohen, L. D. Cohen, and N. Ayache. Introducing new deformable surfaces to segment 3D images. In *Conference on Computer Vision and Pattern Recognition*, pages 738–739, 1991.

[Cryer *et al.*, 1992] J. E. Cryer, Ping-Sing Tsai, and Mubarak Shah. Combining shape from shading and stereo using human vision model. Technical Report CS-TR-92-25, U. Central Florida, 1992.

[Delingette *et al.*, 1991] H. Delingette, M. Hebert, and K. Ikeuchi. Shape representation and image segmentation using deformable surfaces. In *Conference on Computer Vision and Pattern Recognition*, pages 467–472, 1991.

[Ferrie *et al.*, 1992] Frank P. Ferrie, Jean Lagarde, and Peter Whaite. Recovery of volumetric object descriptions from laser rangefinder images. In *European Conference on Computer Vision*, Genoa, Italy, April 1992.

[Fua and Hanson, 1991] P. Fua and A.J. Hanson. An optimization framework for feature extraction. *Machine Vision and Applications*, 4(2):59–87, Spring 1991.

[Fua and Leclerc, 1990] P. Fua and Y.G. Leclerc. Model driven edge detection. *Machine Vision and Applications*, 3:45–56, 1990.

[Fua and Leclerc, 1993] P. Fua and Y.G. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. In *ARPA Image Understanding Workshop*, Washington, D.C., April 1993. Also available as Tech Note 535, Artificial Intelligence Center, SRI International.

[Fua and Sander, 1992] P. Fua and P. Sander. Segmenting unstructured 3d points into surfaces. In *European Conference on Computer Vision*, Genoa, Italy, April 1992.

[Fua, 1993] P. Fua. A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine Vision and Applications*, 6(1), Winter 1993. Available as INRIA research report 1369.

[Heipke, 1992] C. Heipke. Integration of digital image matching and multi image shape from shading. In *International Society for Photogrammetry and Remote Sensing*, pages 832–841, Washington D.C., 1992.

[Kass *et al.*, 1988] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[Leclerc, 1989a] Y. G. Leclerc. *The Local Structure of Image Intensity Discontinuities.* PhD thesis, McGill University, Montréal, Québec, Canada, May 1989.

[Leclerc, 1989b] Y.G. Leclerc. Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, 3(1):73–102, 1989.

[Liedtke *et al.*, 1991] C. E. Liedtke, H. Busch, and R. Koch. Shape adaptation for modelling of 3D objects in natural scenes. In *Conference on Computer Vision and Pattern Recognition*, pages 704–705, 1991.

[Lowe, 1991] D. G. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(441-450), 1991.

[Luenberger, 1984] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, Menlo Park, California, second edition, 1984.

[Ma and Thonnat, 1992] R. Ma and M. Thonnat. A robust and efficient contour-based stereo matching algorithm. Research report (in preparation), INRIA, 1992.

[Meygret *et al.*, 1990] A. Meygret, M. Thonnat, and M. Berthod. A pyramidal stereovision algorithm based on contour chain points. In *European Conference on Computer Vision*, pages 83–88, Antibes, France, April 1990.

[Pentland and Sclaroff, 1991] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:715–729, 1991.

[Pentland, 1990] A. Pentland. Automatic extraction of deformable part models. *International Journal of Computer Vision*, 4(2):107–126, March 1990.

[Press *et al.*, 1986] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes, the Art of Scientific Computing.* Cambridge U. Press, Cambridge, MA, 1986.

[Quam and Strat, 1991] L. Quam and T.M. Strat. SRI image understanding research in cartographic feature extraction. In *International Society for Photogrammetry and Remote Sensing*, Munich, Germany, September 1991. Also available as Tech Note 505, Artificial Intelligence Center, SRI International.

[Rissanen, 1987] J. Rissanen. *Encyclopedia of Statistical Sciences*, volume 5, chapter Minimum-Description-Length Principle, pages 523–527. John Wiley and Sons, New York, New York, 1987.

[Robert and Faugeras, 1991] L. Robert and O.D. Faugeras. Curve-Based Stereo: Figural Continuity and Curvature. In *Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, June 1991.

[Stokely and Wu, 1992] E. M. Stokely and S. Y. Wu. Surface parameterization and curvature measurement of arbitrary 3-d objects: five practical methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):833–839, August 1992.

[Szeliski and Tonnesen, 1992] R. Szeliski and D. Tonnesen. Surface modeling with oriented particle systems. In *Computer Graphics (SIGGRAPH'92)*, pages 185–194, July 1992.

[Terzopoulos and Metaxas, 1991] D. Terzopoulos and D. Metaxas. Dynamic 3D models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(703-714), 1991.

[Terzopoulos and Vasilescu, 1991] D. Terzopoulos and M. Vasilescu. Sampling and reconstruction with adaptive meshes. In *Conference on Computer Vision and Pattern Recognition*, pages 70–75, 1991.

[Terzopoulos, 1986] D. Terzopoulos. Regularization of inverse visual problems involving discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:413–424, 1986.

[Vaillant and Faugeras, 1992] R. Vaillant and O.D. Faugeras. Using Occluding Contours for 3D Object Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, February 1992.

[Vemuri and Malladi, 1991] B. C. Vemuri and R. Malladi. Deformable models: Canonical parameters for surface representation and multiple view integration. In *Conference on Computer Vision and Pattern Recognition*, pages 724–725, 1991.

Appendix B

"Registration without Correspondences"

CVPR; Seattle, Washington, June, 1994
P.V. Fua and Y.G. Leclerc

Registration without Correspondences

P. Fua and Y.G. Leclerc
SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025, USA
(fua@ai.sri.com leclerc@ai.sri.com)

November 23, 1993

Abstract

In this paper, we present a method for registering images of complex 3–D surfaces that does not require explicit correspondences between features across the images. Our method relies on the use of a full 3–D model of the surface. This approach constrains the camera parameters strongly enough so that the models do not need, initially, to be accurate to yield good results. Furthermore, when registration has been achieved, the models can be refined and the fine details recovered.

We use the 3–D surface model to adjust the position and orientation of the camera by minimizing an objective function based on the projections of the images onto the model. The method presented here complements our approach, described in previous publications, to the recovery of 3–D surface models from multiple images whose camera parameters are known.

Our method is applicable to the calibration of stereo imagery, the precise registration of new images of a scene and the tracking of deformable objects. It can therefore lead to important applications in fields such as augmented reality in a medical context or data compression for transmission purposes. We demonstrate its applicability by using both synthetic images and real images of faces and of terrain.

Keywords : Registration, Calibration, Surface reconstruction, Stereo,

1 Introduction

Most of the work in recovering camera position and orientation from a set of images relies on extracting point-like features from these images. Many calibration methods rely on imaging an object whose geometry is known with great precision and exhibits features that are easy to detect; these features and their known 3–D positions can then be used to compute both the external and internal camera parameters. The works described in [Faugeras and Toscani, 1986, Tsai, 1989, Baltsavias, 1991] are examples of this approach. In the area of cartography, the calibration object is replaced by landmarks. Since landmarks or calibration grids are not always available, a large number of methods have been developed to recover relative external camera parameters, [Longuet-Higgins, 1981, Genery, 1779, Weng *et al.*, 1989, Zhang, 1993] for example, and even internal ones [Luong and Faugeras, 1993] without them. They typically use point correspondences between images and are very sensitive to errors in these correspondences, even though eliminating outliers can mitigate the problem [Fischler and Bolles, 1981].

In this paper, we present a method for registering images of objects with complex surfaces without extracting features or generating explicit correspondences. We define registration here as the estimation of the external camera parameters for two or more images given an estimate of the object's shape and assuming that the internal parameters are known. We show that our method is capable of recovering the camera parameters even when the shape of the object is known only approximately.

We use a full three dimensional (3–D) surface model to adjust the position and orientation of the cameras by minimizing an objective function based on the projections of the model into the images. The projections are computed using the current estimate of the external camera parameters and the fixed internal ones. Each point on the 3–D surface potentially projects into more than one image. In the simplest case, the objective function is the sum, over each point on the surface of the 3–D surface, of the variance of the gray levels of the point's projections in the images in which it is visible; a hidden-surface algorithm is used to account for self-occlusions. Standard optimization techniques, such as conjugate gradient descent and the simplex algorithm, are used to minimize the objective function.

The above method complements our previously described approach to the recovery of 3–D surface models from multiple images whose camera parameters are known [Fua and Leclerc, 1993a, Fua and Leclerc, 1993b]. Instead of adjusting only the surface shape to minimize the objective function as in these previous papers, here we adjust the camera parameters (and in some cases, iteratively adjust the camera parameters and surface shape). Other components of the objective function, as described in these papers, could be used in the optimization procedure. However, for the sake of brevity, we describe here only the "stereo" component and refer the interested reader to our previous publications.

Our method has several potential applications. In the simplest case, one can assume that the 3–D model is in fact an accurate representation of the object in the images. For

example, recent work in medical imaging attempts to combine images of 3–D models derived from computer tomography (CT) and/or magnetic resonance imaging (MRI) scans with live images of the patient's shaved anatomy to assist in surgery [Lorensen *et al.*, 1993]. In this case, the 3–D model can be assumed to be accurate, and the problem is to register the 3–D model with the incoming imagery. Our approach is ideal for this situation because it does not depend on having previously defined features manually aligned with the 3–D model.

A more complex situation obtains when neither a precise 3–D model nor accurate camera models are available. For example, in the area of cartography, this may happen when a new image of a site is acquired for which only a rough elevation model (DEM) is available. In this case, we combine our surface recovery and camera parameter recovery methods into an iterative procedure: first estimate the 3–D model, then refine the camera parameters, and repeat until a stable solution is reached. Our approach offers a fully automatic procedure for block adjustment without the use of ground control points or manually designated pass points.

Our method's ability to recover external camera parameters as well as shape using approximate surface models is also applicable to the tracking of deformable objects. For example, a model can be acquired from an initial pair or triplet of images, and can then be used to track the object's motion in subsequent frames while its shape is being iteratively adjusted.

Our iterative approach is described in Section 2 and its stability and convergence properties described in Section 3. Our experimental results, on both synthetic and real data, show that convergence can be achieved for errors in camera parameters resulting in an average absolute deviation of up to 10 pixels in the projection of the surface points into the images relative to the correct answer. Although a larger range of errors in the initial estimate of camera parameters could be accomodated by if a coarse-to-fine strategy were to be used, the purpose of the experiments described here was to determine the range of errors in camera parameters that could be accomodated at any given level using a coarse-to-fine strategy. Hence the use of a dimensionless quantity, pixels, as the measure of error in the camera parameters. In Section 3, we also show that the registrations we derive are good enough for precise surface recovery. Finally, we show an application of our method to the tracking of a person whose facial expression is changing and demonstrate that the shape estimate acquired in one position is good enough to compute both the overall head motion and the deformation in the shape of the face.

2 Meshes

Our approach to recovering surface shape and camera parameters is to deform a 3–D representation of the surface so as to minimize an objective function. The free variables of this objective function are the coordinates of the vertices of the mesh representing the surface and six external parameters for each camera. The process is started with an initial estimate

of the surface and camera parameters.

We represent a surface S by a hexagonally connected set of vertices called a *mesh*. The position of each vertex is specified by its x,y, and z Cartesian coordinates, and each vertex in the interior of the surface has exactly six neighbors. Neighboring vertices are further organized into triangular planar surface elements called *facets*. In Figure 1(a), we show a wireframe representation of such a mesh.

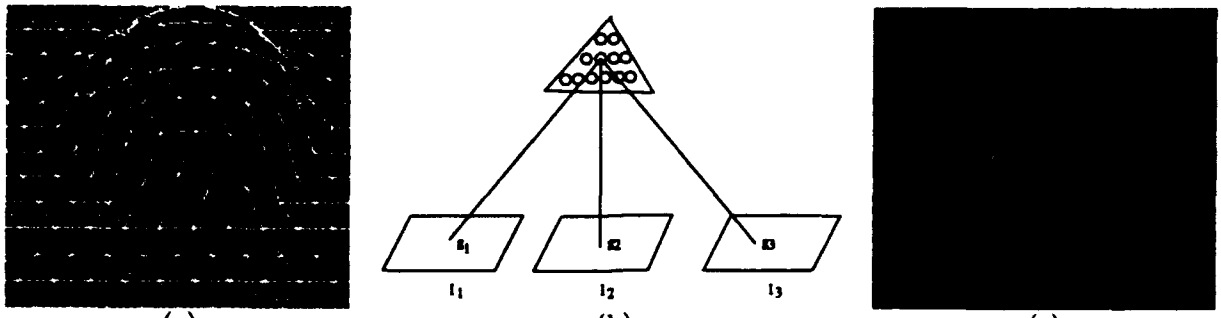


Figure 1: Projection of a mesh, and the Facet-ID image used to accommodate occlusions during surface reconstruction: (a) A wire-frame representation of the mesh (bold white lines) and the sample points in each facet (interior white points). (b) Facets are sampled at regular intervals as illustrated here. The stereo component of the objective function is computed by summing the variance of the gray level of the projections of these sample points, the g_is. (c) The Facet-ID image, wherein the color at a pixel is chosen to uniquely identify the visible facet at that point (shown here as a gray-level image). It is used for visibility computations

Assuming that the internal camera parameters are known, we specify the camera positions as follows. We take the position of the first one as our reference; we then specify the deviations from the initial estimate of the external parameters of the other cameras by defining a vector C with three rotation angles and three translations per camera.

In its most general form [Fua and Leclerc, 1993a, Fua and Leclerc, 1993b], the objective function $\mathcal{E}(S,C)$ that we use to recover the surface is a sum of terms that take into account the image-based constraints—stereo and shape from shading—and the externally supplied geometric constraints—features and silhouettes—that can be brought to bear on the surface. In our registration work, we have so far used only the stereo term, although we plan to use the others in the future. In the remainder of this section we first present our optimization procedure. We then describe in detail the implementation of the stereo energy term and show that it is well adapted to our registration problem.

2.1 Optimization Procedure

In theory, we could simultaneously optimize $\mathcal{S}$ and $\mathcal{C}$. However, in practice we have found it more effective to adjust them sequentially. Given an estimate of the camera parameters, we recover the surface's shape by minimizing $\mathcal{E}(\mathcal{S},\mathcal{C})$ with respect to $\mathcal{S}$. We then use the result to improve the camera parameters by minimizing $\mathcal{E}(\mathcal{S},\mathcal{C})$ with respect to $\mathcal{C}$. In Section 3, we show that this process converges in a few iterations.

We optimize the camera parameters using standard implementations of either the conjugate gradient or simplex algorithms [Press *et al.*, 1986] with very similar results.

To recover the surfaces's shape, because of the non-convexity ot the objective function, we use an optimization method that is inspired by the heuristic technique known as a *continuation method* [Terzopoulos, 1986, Leclerc, 1989a, Leclerc, 1989b] in which we add a regularization term to the objective function and progressively reduce its influence. We define the total energy of the mesh, $\mathcal{E}_T(\mathcal{S})$, as

$$\begin{aligned} \mathcal{E}_T(\mathcal{S}) &= \lambda_D \mathcal{E}_D(\mathcal{S}) + \mathcal{E}(\mathcal{S}) \,, \\ \mathcal{E}_(\mathcal{S}) &= \sum_i \lambda_i \mathcal{E}_i(\mathcal{S}) \,. \end{aligned}$$

The $\mathcal{E}_i(\mathcal{S})$ represent the image- and geometry-based constraints, and the λ_i their relative weights. $\mathcal{E}_D(\mathcal{S})$, the regularization term, serves a dual purpose. First, we define it as a quadratic function of the vertex coordinates, so that it "convexifies" the energy landscape when λ_D is large and improves the convergence properties of the optimization procedure. Second, in the presence of noise, some amount of smoothing is required to prevent the mesh from overfitting the data, and wrinkling the surface excessively. This is especially true when dealing with decalibrated data as shown in Figure 2.

In our implementation [Fua and Leclerc, 1993b], we take $\mathcal{E}_D$ to be a quadratic term that approximates the curvature or local deviation from a plane at every vertex. It is amenable to a "snake-like" optimization technique [Kass *et al.*, 1988] in which every iteration can be reduced to solving a set of sparse linear equations. It can be shown that the dynamics of the optimization are controlled by the gradient of the objective function. As a result, we have found that an effective way to normalize the contributions of the various components of the objective function is to define a set of user-specified weights λ'_i such that

$$\sum_{1 \le i \le n} \lambda'_i < 1 \,.$$

These weights are then used to define the λs as follows:

$$\lambda_i = \frac{\lambda'_i}{\| \vec{\nabla} \mathcal{E}_i(\mathcal{S}^0) \|},$$

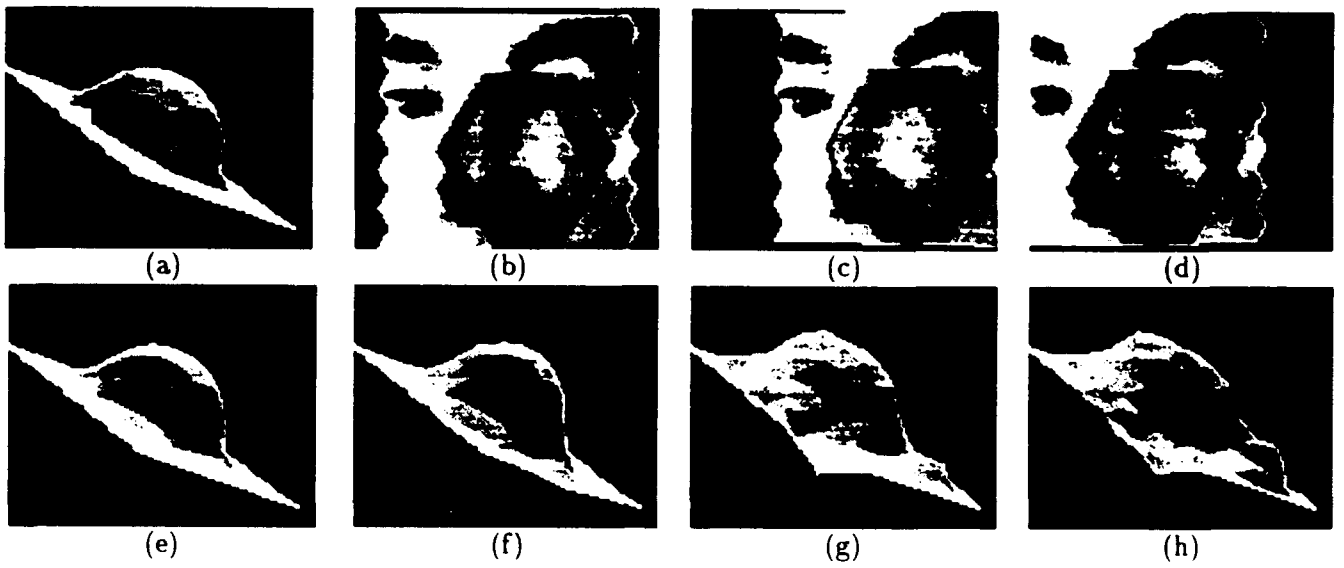


Figure 2: (a) Shaded view of a hemisphere. (b,c,d) Synthetic images generated by texture mapping the image of a face onto the hemisphere. (e,f,g,h) Recovered surface using our surface reconstruction method and progressively worse camera models. The perturbations of the camera models, as defined in Section 3.1, range from 0 to 6 pixels.

$$\lambda_D = \frac{\lambda'_D}{\| \vec{\nabla} \mathcal{E}_D(\mathcal{S}^0) \|},$$

where $\lambda'_D = 1 - \sum_i \lambda'_i$ and $\mathcal{S}^0$ is the surface estimate at the start of each optimization step. Because the normalization makes the influence of the various terms comparable irrespective of actual radiometry or dimensions, we have found that the user-specified λ'_i weights are context-specific but not image-specific. In other words, we may use one set of parameters for images of faces and another when dealing with aerial images, but we do not have to change them for different faces or different landscapes.

The continuation method discussed above is implemented by taking the initial value of λ'_D to be 0.5 and then progressively decreasing it while keeping the relative values of the λ'_is constant.

In this work, while trying to recover both camera parameters and object shape, in addition to the smoothness term, we use only the stereo term of our objective function that we denote $\mathcal{E}_{St}$ and describe below.

2.2 Stereo: Multi-Image Intensity Correlation

The basic premise of most correlation-based stereo algorithms is that the projection of the 3-D points into various images, or at least band-passed or normalized versions of these images, must have identical grey-levels. To take advantage of this property in our object-centered representation, we define the stereo component of our objective function as the variance in gray-level intensity of the projections in the various images of a given sample-point on a facet, summed over all sample-points, and summed over all facets. This component is presented in stages in the remainder of this subsection.

First, we define the sample-points of a facet by noting that all points on a triangular facet are a convex combination of its vertices. Thus, we can define the sample-points $\mathbf{x}_{k,l}$ of facet f_k as

$$\mathbf{x}_{k,l} = \lambda_{l,1}\,\mathbf{x}_{k,1} + \lambda_{l,2}\,\mathbf{x}_{k,2} + \lambda_{l,3}\,\mathbf{x}_{k,3},\ l = 4, \ldots n_s,$$

where $\mathbf{x}_{k,1}$, $\mathbf{x}_{k,2}$, and $\mathbf{x}_{k,3}$ are the coordinates of the vertices of facet f_k, and $\lambda_{l,1}+\lambda_{l,2}+\lambda_{l,3} = 1$. In practice, $\lambda_{l,1}$ and $\lambda_{l,2}$ are both picked at regular intervals in $[0,1]$. When their sum is smaller than one, $\lambda_{l,3}$ is taken to be $1 - \lambda_{l,1} - \lambda_{l,2}$. In Figure 1(b), we see an example of the sample-points of a facet.

Next, we develop the sum of squared differences in intensity from all images for a given point $\mathbf{x}$. A point $\mathbf{x}$ in space is projected into a point $\mathbf{u}$ in image g_i via the perspective transformation $\mathbf{u} = \mathbf{m}_i(\mathbf{x})$. Consequently, the sum of squared differences in intensity from all the images, $\sigma'^2(\mathbf{x})$, is defined by

$$\begin{aligned} \mu'(\mathbf{x}) &= \frac{1}{n_i}\sum_{i=1}^{n_i} g_i(\mathbf{m}_i(\mathbf{x}))\,, \\ \sigma'^2(\mathbf{x}) &= \frac{1}{n_i}\sum_{i=1}^{n_i} (g_i(\mathbf{m}_i(\mathbf{x})) - \mu'(\mathbf{x}))^2\,. \end{aligned}$$

Figure 1(b) illustrates the projection of a sample-point of a facet onto several images. The above definition of $\sigma'^2(\mathbf{x})$ does not take into account occlusions of the surface. To do so, we use a "Facet-ID" image, shown in Figure 1(c). It is generated by encoding the index i of each facet f_i as a unique color and projecting the surface into the image plane, using a standard hidden-surface algorithm.[1] Thus, when a sample-point from facet f_k is projected into an image, the index k is compared to the index stored in the Facet-ID image at that point. If they are the same, then the sample-point is visible in that image; otherwise, it is not. Let $v_i(\mathbf{x}) = 1$ when point $\mathbf{x}$ is determined to be visible in image g_i by the method

[1] Our algorithm is implemented on Silicon Graphics machines whose graphics hardware allows for fast computation of the Facet-ID image.

above, and $v_i(\mathbf{x}) = 0$ otherwise. Then, the correct form for the sum of squared differences in intensity at a point $\mathbf{x}$ is defined by

$$\mu(\mathbf{x}) = \frac{\sum_{i=1}^{n_i} v_i(\mathbf{x}) g_i(\mathbf{m}_i(\mathbf{x}))}{\sum_{i=1}^{n_i} v_i(\mathbf{x})},$$

$$\sigma^2(\mathbf{x}) = \frac{\sum_{i=1}^{n_i} v_i(\mathbf{x}) \left(g_i(\mathbf{m}_i(\mathbf{x})) - \mu(\mathbf{x})\right)^2}{\sum_{i=1}^{n_i} v_i(\mathbf{x})}.$$

When the sample-point is visible in fewer than two images (that is, when $\sum_{i=1}^{n_i} v_i(\mathbf{x}) < 2$), the above variance has no meaning and is taken to be 0. Let s_k denote the number of facet samples for facet k for which the variance is meaningful. Summing $\sigma^2(\mathbf{x})$ over all sample-points and over all facets and normalizing by the number of meaningful sample-points yields the multi-image intensity correlation component $\mathcal{E}_{St}$:

$$\mathcal{E}_{St}(\mathcal{S}) = \frac{\sum_{k=1}^{n_f} \sum_{l=4}^{n_s} \sigma^2(\mathbf{x}_{k,l})}{\sum_{k=1}^{n_f} s_k}.$$

In Figure 3 we show the result of running the stereo component of our objective function on an aerial stereo pair of a sharp ridge. We start with a coarse DEM that was provided to us by the U.S. Geological Survey (USGS) and refine it using our method. Note that the recovered ridge, although it is seen almost edge-on, has become much sharper than in the original model result and that details in the upper part of the image are well recovered. In Figure 4, we show the reconstruction of a face using a triplet of images and adding to the stereo term the shape-from-shading term described in a previous publication [Fua and Leclerc, 1993a]. For both scenes, precise camera models have been computed using resection in one case and the INRIA calibration set-up in the other [Faugeras and Toscani, 1986]. In Section 3, we perturb these camera models to study the sensitivity of our method to errors in the camera models and its ability to remove them.

Our formulation is well adapted to tackling the registration problem because:

- It is object-centered and incorporates the use of camera models. It can therefore naturally be used to optimize both the 3–D coordinates of the surface vertices and the camera parameters.
- It can model surfaces that are slanted sharply away from the cameras. It avoids the constant depth within the window assumption of simple correlation algorithms and can therefore deal with surfaces of arbitrary orientation.
- It can deal with an arbitrary number of images, thereby strongly constraining the problem.

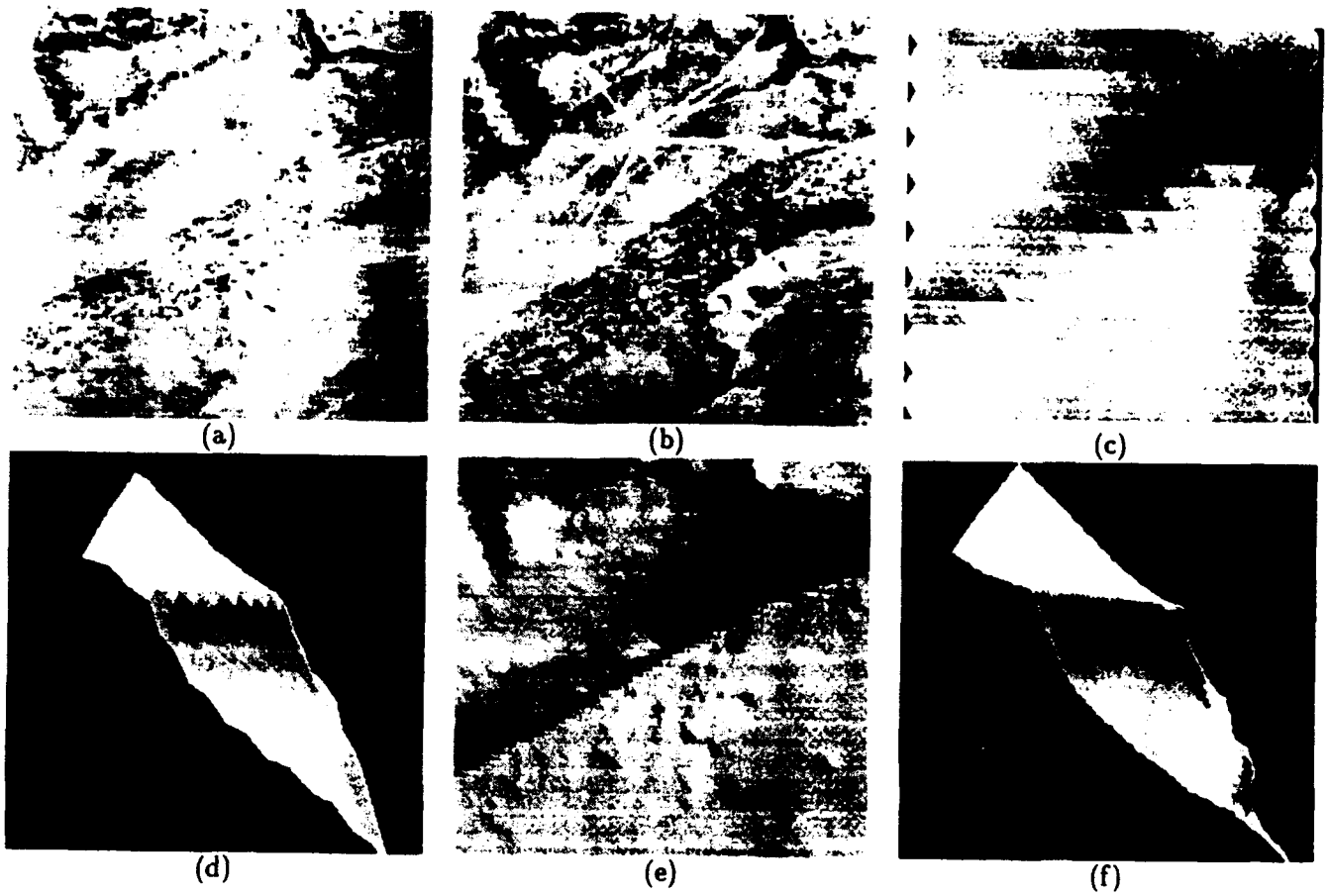


Figure 3: (a,b) A stereo pair of aerial images of a hilly site. In (b), the arrows point toward the sharp ridge in the center of the images and the shadow-casting cliffs at the top. (c) Shaded view of the triangulated DEM, seen from the viewpoint of (a). (d) Shaded view of the DEM as seen by an observer located above the upper left corner of the scene. (e,f) Shaded views of the mesh after optimization, as seen from the viewpoints of (c) and (d). Note that the recovered ridge has become very sharp and that the shadow casting cliffs are clearly visible at the top of (e) and the bottom right corner of (f).

- It can correctly handle the self-occlusions that inevitably arise when dealing with complex surfaces.

We now turn to our experimental results.

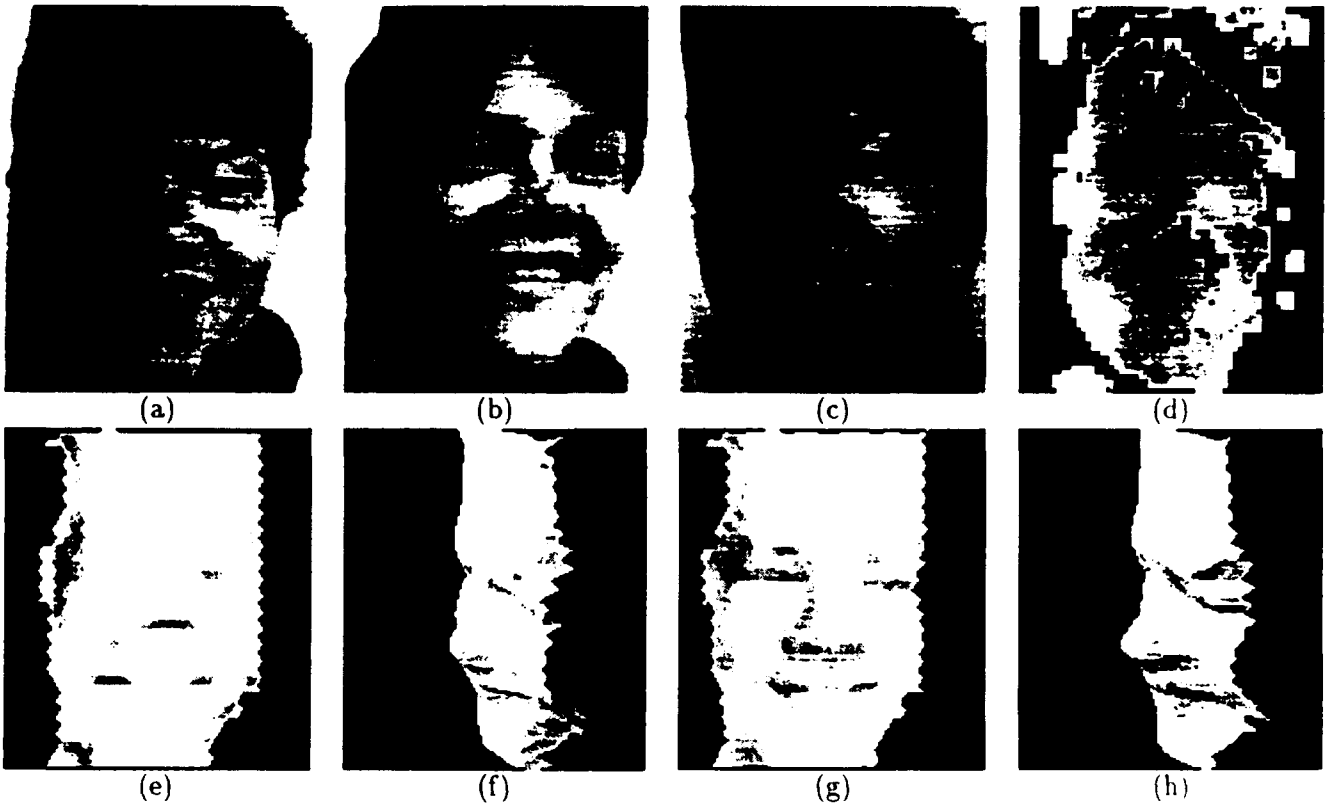


Figure 4: (a,b,c) A triplet of face images with known camera models (courtesy of INRIA) (d) Disparity map computed using a correlation-based algorithm. The black areas indicate that the stereo algorithm could not find a match. Elsewhere, lighter grays indicate greater distances from the image plane. (e,f) Initial surface estimate derived by smoothing and interpolating the disparity map and shown as a shaded surface seen from two different viewpoints. (g,h) Final surface estimate derived by combining stereo and shape-from-shading, as described in [Fua and Leclerc, 1993a].

3 Experimental Results

In this section we first quantify the ability of our method to recover external camera parameters by using the synthetic images of Figure 2, the aerial images of Figure 3 and the face images of Figure 4. We use the images'known camera models as our reference. We demonstrate the robustness of our method by starting with a rough surface model, substantially perturbing the camera models and showing that registration can be achieved nevertheless.

We then show that both surface and camera parameters can be adequately recovered in the more difficult situation where the camera parameters are known only approximately and no rough estimate of the surface is initially available. Finally, we use different sets of face images to show possible applications of this technique to motion tracking and 3–D graphics.

3.1 Sensitivity Analysis

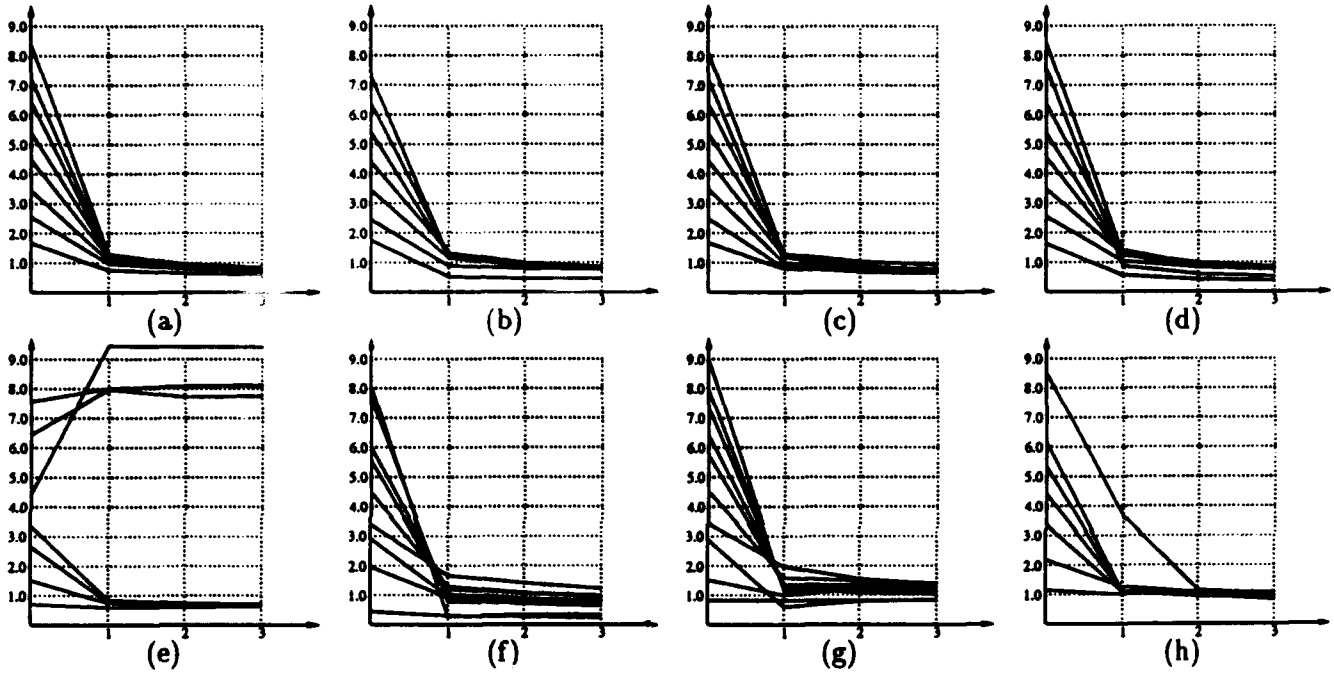


Figure 5: Perturbing and then recovering camera parameters. The meaning of the eight curves is explained in detail in Section 3.1.

Here we study the sensitivity of our method to decalibration. For each of our test scenes, we perturb the positions and orientations of the second and third cameras by random amounts. The perturbation's magnitude is defined as the average absolute deviation of the projection of the mesh vertices, measured in pixels. The eight graphs of Figure 5 depict the results. Each of these graphs summarizes the result of one hundred trials, where a trial consists of the following steps:

- Perturb the camera models by multiplying the corresponding 3x4 projection matrix by a 4x4 translation and rotation matrix derived from six random translation and rotation values.

- Optimize the z coordinates of the mesh using the perturbed camera models and a fixed value of the regularization parameter λ'_D defined in Section 2.1.
- Optimize the camera orientation parameters using the deformed mesh.
- Reoptimize the z and external parameters twice more.

For each trial, we have recorded the initial camera perturbation and the residual perturbation after each optimization of the camera parameters. For each graph, we have grouped the trials according to the magnitude of the initial camera perturbation—between 0 and 1 pixels, 1 and 2 pixels, ...,8 and 9 pixels—and averaged the values in each group resulting in the various plots. Ideally, all the curves should converge to zero.

The four graphs of the first row of Figure 5 were generated using the three synthetic images of Figure 2, a nosiy hemisphere as our initial surface estimate and four different values of the λ'_D regularization parameters, 0.5, 0.6, 0.7 and 0.8. Note that the curves are fairly similar in shape and converge to deviation values between 0.5 pixel and 1.0 pixel. In other words, the regularization term introduces a small bias but the method does not appear to be very sensitive to the exact amount of smoothing. We are currently investigating the use of a coarse-to-fine strategy in mesh resolution to try to improve upon this result.

The second row of Figure 5 depicts the result obtained using real images. Graph(e) corresponds to the aerial images using the approximative DEM shown in Figure 3(d) as our initial estimate. The other three graphs were generated using the face images of Figure 4 and the rough surface model depicted in Figure 4(e). For graph(f), we used only the first two images of the triplet, for graph(g) the first and third images of the triplet, and for graph(h) all three images simultaneously. The overall behavior of the curves is the same as in the synthetic case.

In the case of the aerial images, however, convergence only happens for smaller values of the initial perturbation. This is not surprising in light of the results of more standard methods for computing camera parameters. It takes six matches of non-coplanar points between two images to compute the external parameters [D.H. Ballard, 1982]. Each vertex gives us a correspondence, but, in this particular case, the terrain is almost flat, except for the sharp ridge in the middle of the image. Excessive perturbation of the camera parameters is tantamount to blurring the ridge, thereby making registration based on surface point correspondences alone underconstrained.

In the case of the faces, the surface is complex enough to allow the recovery of camera orientation parameters, even for fairly large initial perturbations. Furthermore, when using all three images at the same time we constrain the problem more strongly and decrease the variability of the results. The curves converge towards deviation values that are slightly higher than those observed in the synthetic case. This can be accounted for by the fact that the camera models we use as a reference are good but not necessarily perfect, they

themselves are not precise to more than 0.5 pixels. In any case, as shown in Section 3.2, the recovered registration has proved sufficient for surface reconstruction purposes.

3.2 Shape and Camera Position Recovery

Here, we demonstrate that the results shown above actually are good enough for accurate surface recovery "from scratch" when the initial camera-models are only approximately known.

We first use the same two scenes as above but use perturbed camera models to initialize our meshes. That is, we use the epipolar lines predicted by the perturbed camera models—they are off by about four pixels—to compute the disparity maps and initial estimates shown in Figure 6(b) and Figure 7(a,b). Because of the decalibration, these initial estimates are poorer than the ones shown in Figures 3 and 4, where we used well-calibrated cameras. Nevertheless, by performing the optimization of both the mesh positions and the camera parameters, we recover the surfaces shown in Figure 6(d,e) and Figure 7(c,d) that are almost indistinguishable from the ones we derived using the well-calibrated data.

So far, we have shown that our method can recover camera position and orientation using approximate surface models. This capability can be used to track deformable objects, as illustrated by Figure 8, which shows two triplets of images of the same face taken at two different times. Note that the person's facial expression has slightly changed. We use our method to detect both the global head motion between the two frames and the local deformation, as follows. We first use our standard method to recover the face in the first triplet, as shown in Figure 8(g). We then find the rotation and translation parameters that minimize our stereo objective function using the first images of both triplets and the previously derived surface estimate. Finally, we use these parameters to rotate and translate all three camera models of the second triplet and use them and the corresponding images to deform the mesh again to better conform to the new facial expression. In Figure 8(h), we show the new estimate of the surface after rotation and translation by the amount computed earlier. In Figure 8(i), we show the displacement of the vertices as a flow field from the old to the new vertex positions. Note that this flow field does not exhibit any global motion, thereby indicating that our global motion estimate is correct. This method could be generalized to the tracking of deformable objects in a video frame while updating of the surface estimate. By using such an approach, the surface model can then be modified incrementally—as opposed to being recomputed at every iteration—and, since we use a physically meaningful 3-D representation, powerful methods such as Kalman filtering can be brought to bear. This technique could be applied to stabilize and transmit face images: The motion parameters would be used to perform the stabilization and one would then only transmit the parts of the face that have undergone a substantial deformation.

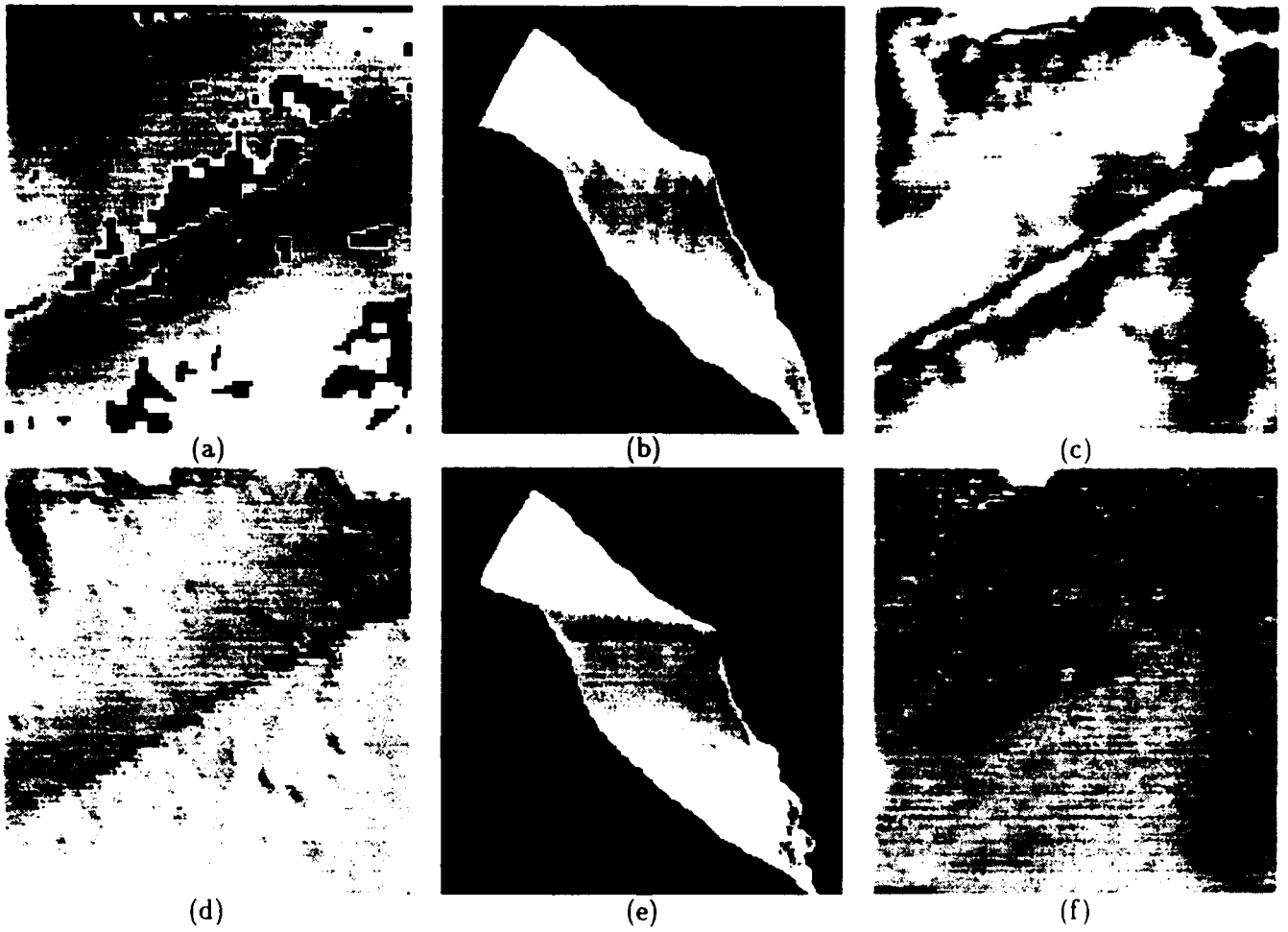


Figure 6: Recalibrating the aerial images of Figure 3. (a) Disparity map computed without camera models by assuming that the epipolar lines are horizontal. (b) Initial surface estimate derived using the disparity map and perturbed camera models. (c) Absolute differences in elevation predicted by this initial model and the optimized model depicted in Figure 3 (e,f). The image is stretched so that differences of more than 50 feet, or about two pixels in disparity, appear in white. (d,e) Shaded views of the mesh after camera parameters recovery and optimization of the mesh following the schedule used to derive the model of Figure 3. (f) Differences in elevation between the two surface estimates. Except for a small patch in the upper part of the image, they are smaller than 25 feet or one pixel in disparity.

Figure 7: Recalibrating the face images of Figure 4. (a,b) Initial surface estimate derived by perturbing the camera models, recomputing a disparity map and interpolating it. Note that the mose is more flattened than previously. (c,d) Final surface estimate after registration and optimization of the initial estimates. They are almost indistinguishable from those of Figure 4.

4 Conclusion

We have presented a method for registering images of complex 3–D surfaces that does not require explicit correspondences between point-like features across the images. Our method relies on the use of a full 3–D model of the imaged surface to recover external camera parameters. This approach constrains the camera parameters strongly enough so that the 3–D models do not need, initially, to be accurate to yield good results. Furthermore, when registration has been achieved, the models can be refined and the fine details in the surfaces of interest recovered precisely.

The method is applicable to the calibration of stereo imagery, the precise registration of new images of a scene, and the tracking of deformable objects. It can therefore lead to important applications in fields such as augmented reality in a medical context or data compression for transmission purposes.

Using static imagery, we have shown that, if the surfaces to be registered have enough relief, the method is both robust and accurate to within 1 pixel for initial errors of up to 10 pixels in camera registration. Future work will concentrate on designing a coarse-to-fine strategy to be able to handle larger errors—ideally, such as those produced by a completely decalibrated set of cameras—and on implementing a Kalman filtering style approach to the modeling of surfaces in video sequences.

Acknowledgments

Support for this research was provided by various contracts from the Advanced Research Projects Agency. We wish to thank Marty Fischler and Olivier Monga whose suggestions have greatly influenced the work described here and Hervé Matthieu for providing us with the face images and corresponding calibration data that have proved extremely valuable to our research effort.

Figure 8: Application to motion tracking of deformable objects. (a,b,c) Triplets of face images taken simultaneously. (courtesy of INRIA). (d,e,f) Second triplet taken slightly later. Note that the head has moved and that the expression has changed. (g) Reconstructed surface for the first triplet. (h) Reconstructed surface for the second triplet, shown rotated and translated so that if matches the first one. (i) Flow field of the motion of the vertices of the first mesh to the rotated second one. Note that the field exhibits no overall structure and only local deformations, showing that the global motion has been correctly recovered.

References

[Baltsavias, 1991] E. P. Baltsavias. *Multiphoto Geometrically Constrained Matching.* PhD thesis, Institute for Geodesy and Photgrammetry, ETH Zurich, December 1991.

[D.H. Ballard, 1982] C.M. Brown D.H. Ballard. *Computer Vision.* Prentice-Hall, 1982.

[Faugeras and Toscani, 1986] O.D. Faugeras and G. Toscani. The calibration problem for stereo. In *Conference on Computer Vision and Pattern Recognition*, pages 15–20, Miami Beach, Florida, 1986.

[Fischler and Bolles, 1981] M.A Fischler and R.C Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications ACM*, 24(6):381–395, 1981.

[Fua and Leclerc, 1993a] P. Fua and Y.G. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. In *ARPA Image Understanding Workshop*, Washington, D.C., April 1993. Also available as Tech Note 535, Artificial Intelligence Center, SRI International.

[Fua and Leclerc, 1993b] P. Fua and Y.G. Leclerc. Using 3–dimensional meshes to combine image-based and geometry-based constraints. Technical Note 536, SRI, October 1993. Submitted to ECCV94.

[Genery, 1779] D.B. Genery. Stereo-camera calibration. In *ARPA Image Understanding Workshop*, pages 101–107, 1779.

[Kass *et al.*, 1988] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[Leclerc, 1989a] Y. G. Leclerc. *The Local Structure of Image Intensity Discontinuities.* PhD thesis, McGill University, Montréal, Québec, Canada, May 1989.

[Leclerc, 1989b] Y.G. Leclerc. Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, 3(1):73–102, 1989.

[Longuet-Higgins, 1981] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[Lorensen *et al.*, 1993] W. Lorensen, H. Kline, C. Nafis, R. Kikinis, D. Altobelli, and L. Gleason. Enhancing reality in the operating room. In *IEEE Visualization Conference*, pages 410–415, San Jose, California, October 1993.

[Luong and Faugeras, 1993] Q.-T. Luong and O.D. Faugeras. Self-calibration of a stereo rig from unknown camera motions and point correspondences. In A. Grun and T.S. Huang, editors, *Calibration and orientation of cameras in computer vision.* Springer-Verlag, 1993. To appear. Also presented at XVII ISPRS, Washington, and INRIA Tech Report RR-2014.

[Press *et al.*, 1986] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes, the Art of Scientific Computing.* Cambridge U. Press, Cambridge, MA, 1986.

[Terzopoulos, 1986] D. Terzopoulos. Regularization of inverse visual problems involving discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:413–424, 1986.

[Tsai, 1989] R.Y. Tsai. Synopsis of Recent Progress on Camera Calibration for 3D Machine Vision. In Oussama Khatib, John J. Craig, and Tomás Lozano-Pérez, editors, *The Robotics Review*, pages 147–159. MIT Press, 1989.

[Weng *et al.*, 1989] J. Weng, T.S. Huang, and N. Ahuja. Motion and structure from two perspective views: algorithms, error analysis and error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(5):451–476, 1989.

[Zhang, 1993] Z. Zhang. Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. In *Conference on Computer Vision and Pattern Recognition*, pages 556–561, 1993.

BIBLIOGRAPHY
(Publications describing work performed on this project)

Y.G. Leclerc and M.A. Fischler, "An optimization based approach to the interpretation of single line drawings as 3-D wire frames," IJCV 9(2):113-136, November, 1992.

P.V. Fua and Y.G. Leclerc, "Combining stereo, shading, and geometric constraints for surface reconstruction from multiple views," Technical Conference Geometric Methods in Computer Vision II of SPIE Symposium, San Diego, California, July, 1993.

P.V. Fua and Y. G. Leclerc, "Object-Centered Surface Reconstruction: Combining Multi-Image Stereo and Shading," IJCV, 1994.

P.V. Fua and Y.G. Leclerc, "Using 3-Dimensional Meshes to Combine Image-Based and Geometry-Based Constraints," ECCV, Stockholm, Sweden, May, 1994.

P.V. Fua and Y.G. Leclerc, "Registration without Correspondences," CVPR, Seattle, Washington, June, 1994.

M.A. Fischler and H.C. Wolf, "Saliency detection and partitioning planar curves," Proc. Image Understanding Workshop, Washington D.C., pp. 917-931, April, 1993.

M.A. Fischler and H.C. Wolf, "Locating perceptually salient points on planar curves," IEEE-PAMI, vol. 16(2):1-17, February, 1994.

W. Neuenschwander, P.V Fua, G. Szekely, and O. Kubler, "Initializing Snakes," CVPR, Seattle, Washington, June, 1994.

W. Neuenschwander, P.V Fua, G. Szekely, and O. Kubler, "Using Boundary Conditions to Improve Snake Convergence," ICPR, Jerusalem, Israel, October, 1994.